Week 1 Core Module 3 IP Project DS 11

Angelo Sang

11/19/2021

ONLINE CRYPTOGRAPHY COURSE ANALYSIS

Business Understanding

Cryptography course was created by a Kenyan entrepreneur 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 dat on the process. Using R programming language we would help her identify which individuals are most likely to click on her ads.

Specifying the Data Analytic Question

Use R Prigramming to identify which individuals are the most likely to click on the ads.

Defining the Metric for Success

We identify the most important features for identifying which individuals are most likely to click on the ads.

Recording the Experimental Design

1) Finding and dealing with outliers, anomalies and missing data within the dataset.
2) Performing univariate and Bivariate analysis

Data Relevance

a) How accurate is the data at identifying which individuals are most likely to click on her ads?
b) Was the dataset sufficient?
c) Was the data biased?
d) Is the data source a reliable source?

```
# loading the dataset and library
library(data.table)
df <- fread('http://bit.ly/IPAdvertisingData')

#previewing the dataset
print(head(df))
```

```
##    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
```

```
print(tail(df))
```

```
##    Daily Time Spent on Site Age Area Income Daily Internet Usage
## 1:                    43.70  28    63126.96               173.01
## 2:                    72.97  30    71384.57               208.58
## 3:                    51.30  45    67782.17               134.42
## 4:                    51.63  51    42415.72               120.37
## 5:                    55.55  19    41920.79               187.95
## 6:                    45.01  26    29875.80               178.35
##                          Ad Topic Line          City Male
## 1:        Front-line bifurcated ability  Nicholasland    0
## 2:        Fundamental modular algorithm     Duffystad    1
## 3:      Grass-roots cohesive monitoring   New Darlene    1
## 4:         Expanded intangible solution South Jessica    1
## 5: Proactive bandwidth-monitored policy   West Steven    0
## 6:      Virtual 5thgeneration emulation   Ronniemouth    0
##                   Country           Timestamp Clicked on Ad
## 1:               Mayotte 2016-04-04 03:57:48             1
## 2:               Lebanon 2016-02-11 21:49:00             1
## 3: Bosnia and Herzegovina 2016-04-22 02:07:01             1
## 4:              Mongolia 2016-02-01 17:24:57             1
## 5:             Guatemala 2016-03-24 02:35:54             0
## 6:                Brazil 2016-06-03 21:43:21             1
```

```
# checking data types
str(df)
```

```
## Classes 'data.table' and 'data.frame':   1000 obs. of  10 variables:
##  $ Daily Time Spent on Site: num  69 80.2 69.5 74.2 68.4 ...
##  $ Age                     : int  35 31 26 29 35 23 33 48 30 20 ...
##  $ Area Income             : num  61834 68442 59786 54806 73890 ...
```

```
##  $ Daily Internet Usage    : num  256 194 236 246 226 ...
##  $ Ad Topic Line           : chr  "Cloned 5thgeneration orchestration" "Monitored national 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               : POSIXct, format: "2016-03-27 00:53:11" "2016-04-04 01:39:02" ...
##  $ Clicked on Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Cleaning the Dataset

```
#Changing the timestamp datatype from factor to date_time
df$Timestamp <- as.Date(df$Timestamp, format = "%Y-%m-%s-%h-%m-%s")

# checking the new datatype for timestamp column
sapply(df, class)
```

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

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

```
## Daily Time Spent on Site                     Age             Area Income
##                        0                       0                       0
##     Daily Internet Usage           Ad Topic Line                    City
##                        0                       0                       0
##                     Male                 Country               Timestamp
##                        0                       0                       0
##            Clicked on Ad
##                        0
```

```
#the data doesn't have missing data
```

```
#checking the columns of the dataframe
colnames(df)
```

```
##  [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"
```

```r
#Checking shape of dataset
dim(df)
```

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

```r
# the dataset has 1000 rows and 10 columns
```

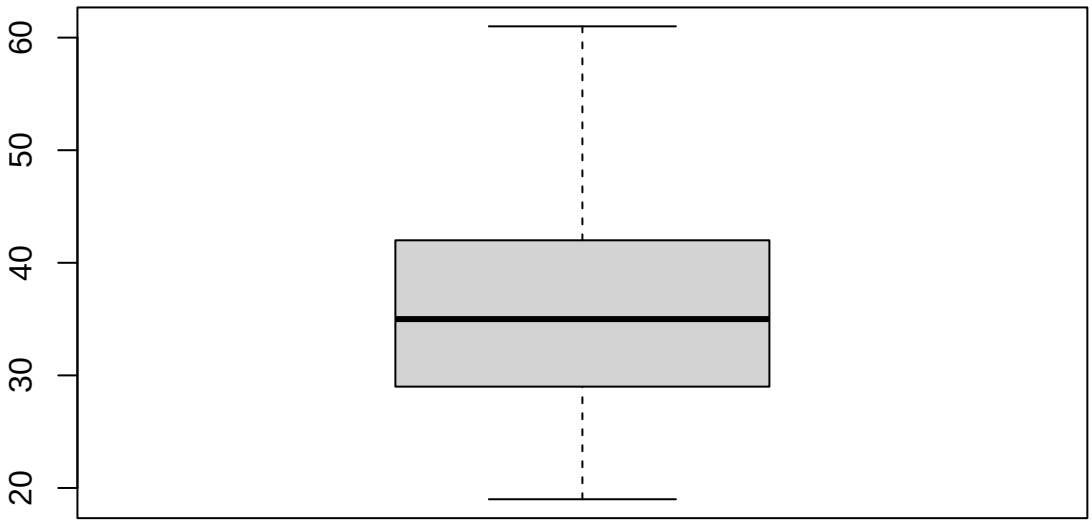
```r
# checking for duplicates
#duplicated(df)
df[duplicated(df)]
```

```
## Empty data.table (0 rows and 10 cols): Daily Time Spent on Site,Age,Area Income,Daily Internet Usage
```

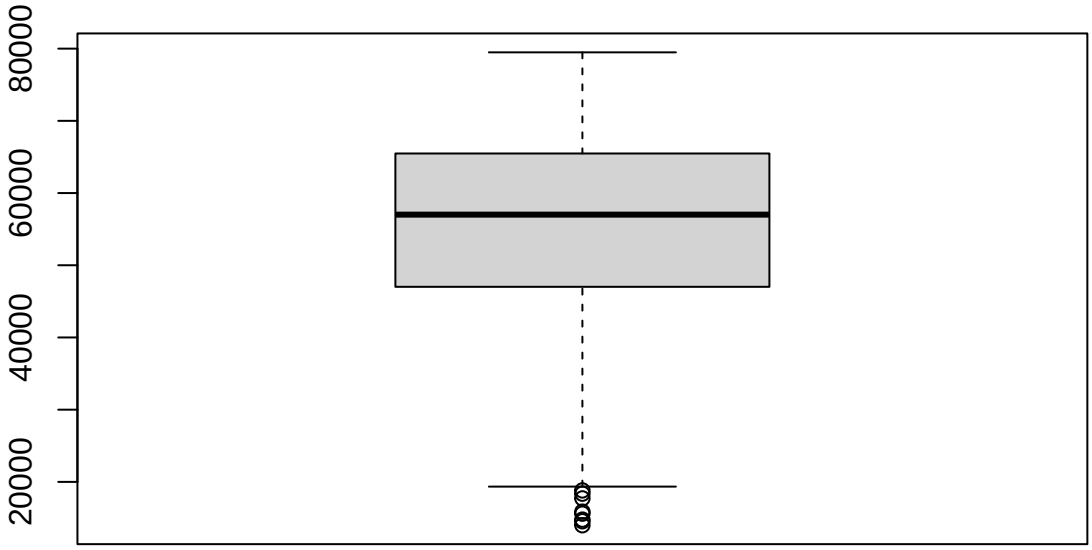
```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

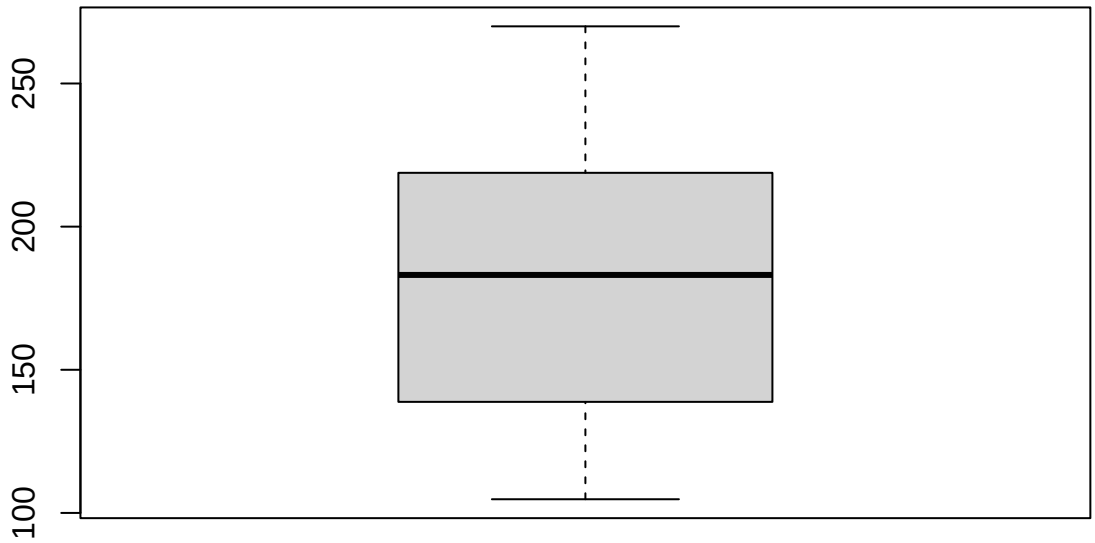
```r
# checking for outliers
boxplot(df$Age)
```




```
boxplot(df$`Area Income`)
```




```
boxplot(df$`Daily Internet Usage`)
```




```
# will not eliminate the outlies because the don't impact our analysis negatively
```

UNIVARIATE ANALYSIS

Age

```
#measures of central tendancy
mean(df$Age)
```

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

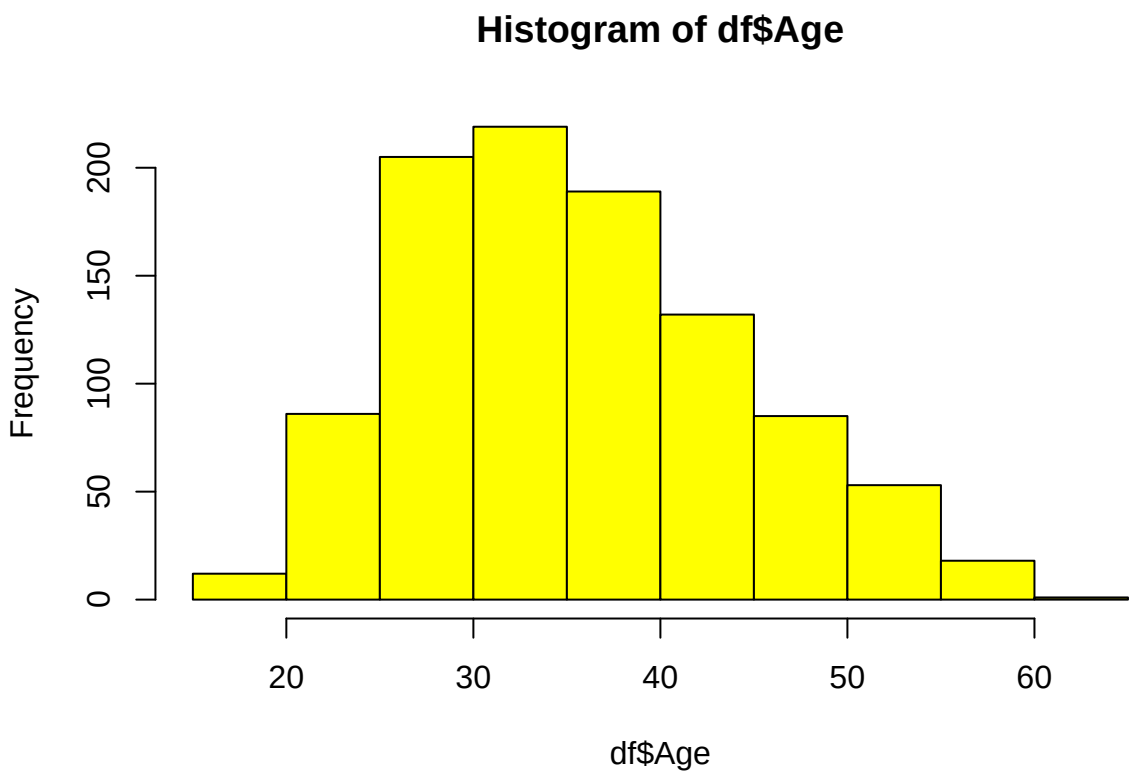
```
median(df$Age)
```

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

```
#mode
age_x <- df$Age
#sort(age_x)
names(table(age_x))[table(age_x)==max(table(age_x))]
```

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

```r
# visualizing the age distribution using histogram
hist(df$Age, col = c('yellow'))
```



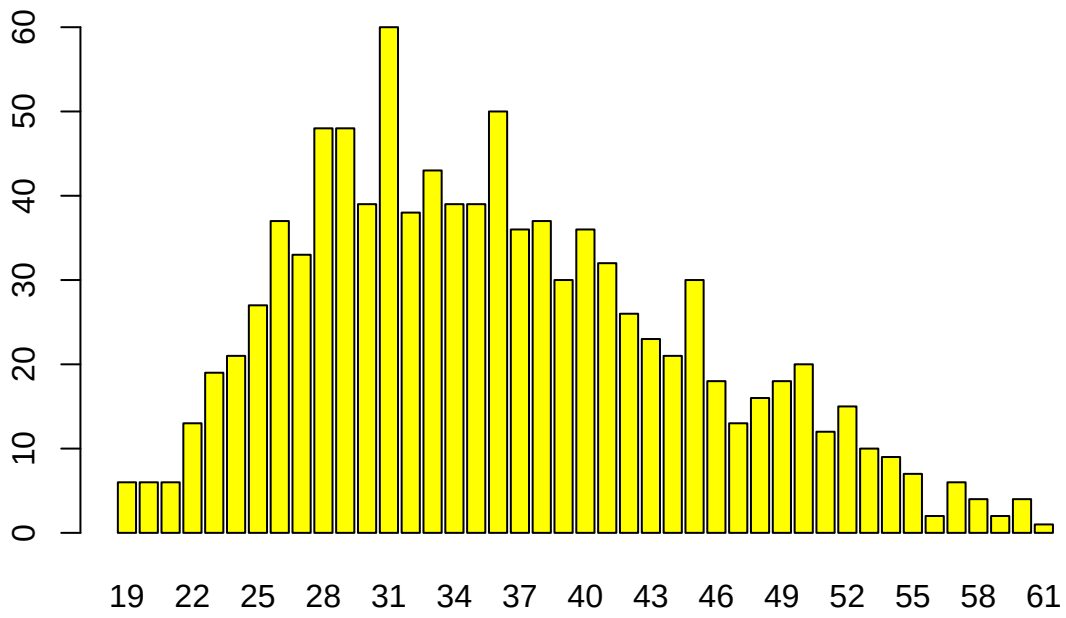

```r
# there is a higher tendency that a person aged between 30 and 35 years of age accessed her ads.
```

- The age distribution is right skewed
- The respondents on the website are mostly 25-40 years old
- The mean age is 36
- The median age is 35

```r
# visualizing the age distribution using bar graph
# fetching the age column
age <- df$Age

# applying the table() function the compute the frequency of the Age column
age_frequency <- table(age)
age_frequency
```

```
## age
## 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
##  6  6  6 13 19 21 27 37 33 48 48 39 60 38 43 39 39 50 36 37 30 36 32 26 23 21
## 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61
## 30 18 13 16 18 20 12 15 10  9  7  2  6  4  2  4  1
```

```
barplot(age_frequency, col = 'yellow')
```




```
# most people accessing her ads where aged 31 years
```

Area_income

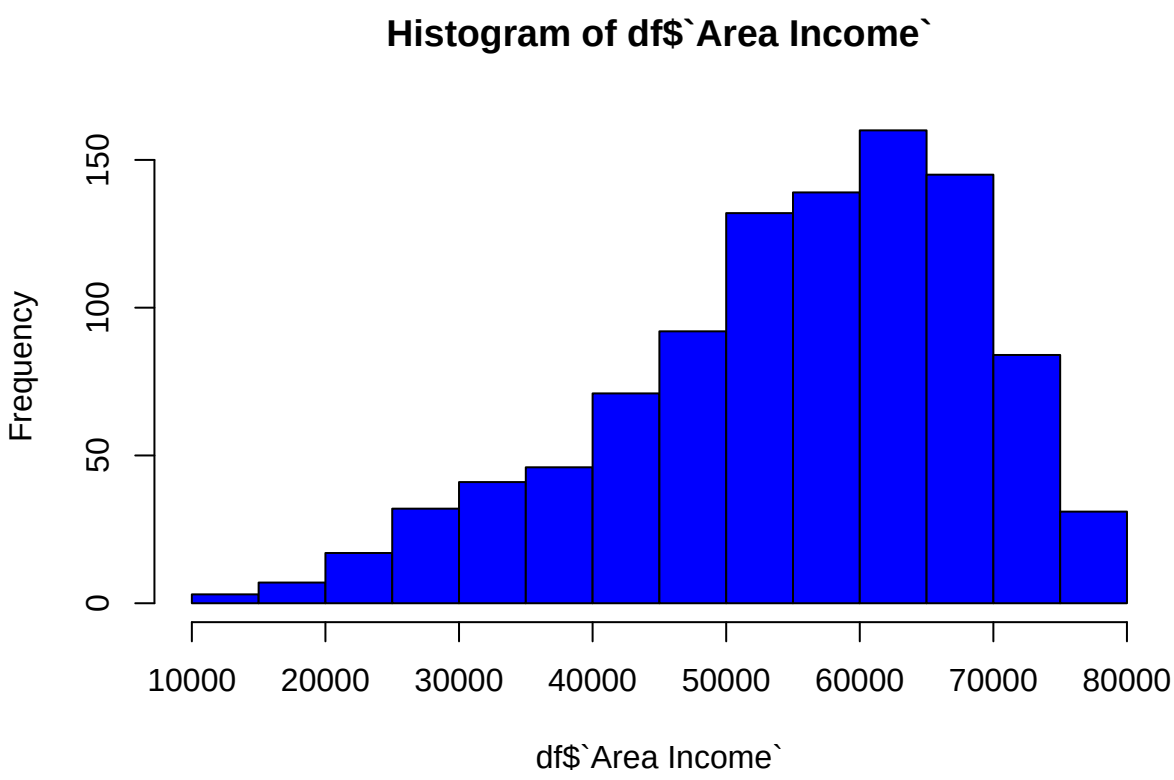
```
#income
#central tendancy
mean(df$`Area Income`)
```

```
## [1] 55000
```

```
median(df$`Area Income`)
```

```
## [1] 57012.3
```

```
# visualizing the area_income distribution using histogram
hist(df$`Area Income`, col = 'blue')
```




```
#people leaving within an area_income between 60,000 and 65,000 had access on her ads the most
```

- The income distribution is Left skewed
- The respondents on the website mostly earn between 55,000 to 70,000
- The mean income is 55,000
- The median income is 57,012

Daily_Internet_Usage

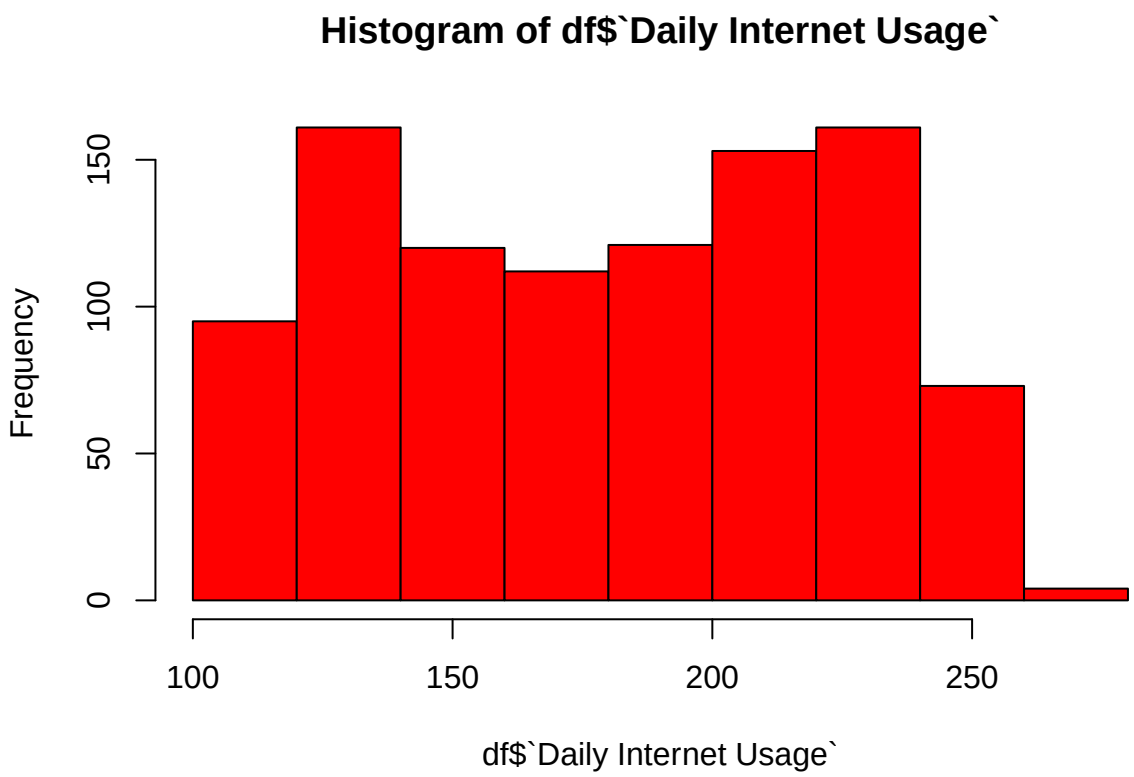
```
#this column represents the amount of data that the user consumes in a day
# measures of central tendency
mean(df$`Daily Internet Usage`)
```

```
## [1] 180.0001
```

```
median(df$`Daily Internet Usage`)
```

```
## [1] 183.13
```

```
# visualizing the daily_internet_usage distribution using histogram
hist(df$`Daily Internet Usage`, col = 'red')
```



- The mean data usage is 180 units - The median data usage is 183.13 units

Ad_Topic_Line

```
ad_topic_line <- df$`Ad Topic Line`

levels(unique(ad_topic_line))
```

```
## NULL
```

```
#factor(unique(ad_topic_line))
```

- All the values are unique in this column thus we would drop it when modelling since it does not provide any additional meaningful information

#City

```
# City where the user is located
# measure of central tendency
length(levels(df$City))
```

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

```
#mode (the modal cities in the dataset)
city_x <- df$City

#sort(city_x). This code gives an ordered list of all the elements in the cities column
names(table(city_x))[table(city_x)==max(table(city_x))]
```

```
## [1] "Lisamouth"    "Williamsport"
```

- The most popular cities in the dataset are: “Lisamouth” “Williamsport”

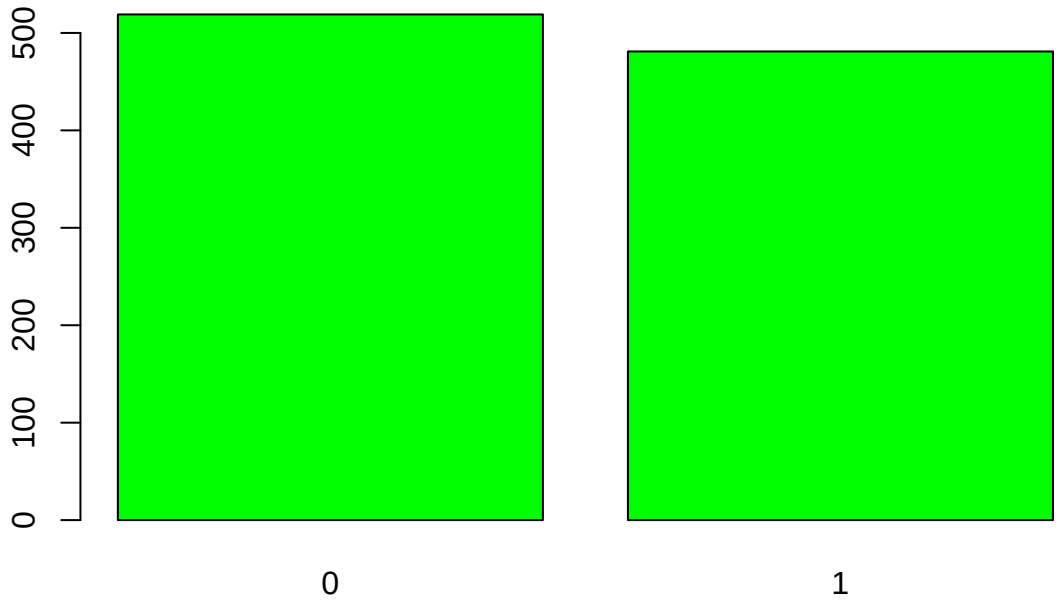
#Gender

```
# gender of the user
# fetching the Male column
gender <- df$Male

# applying the table() function to compute the frequency of the Male column
gender_frequency <- table(gender)
gender_frequency
```

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

```
# 1 represented the males and 0 represented the females
```

```
# visualizing the Gender distribution using bar graph
barplot(gender_frequency, col = 'green')
```




```
# most of the people who accessed her ads were females
```

#Daily_Time_Spent_on_Site

```
# This column represents the amount of time that the user spends on the website
# measures of central of tendency
mean(df$`Daily Time Spent on Site`)
```

```
## [1] 65.0002
```

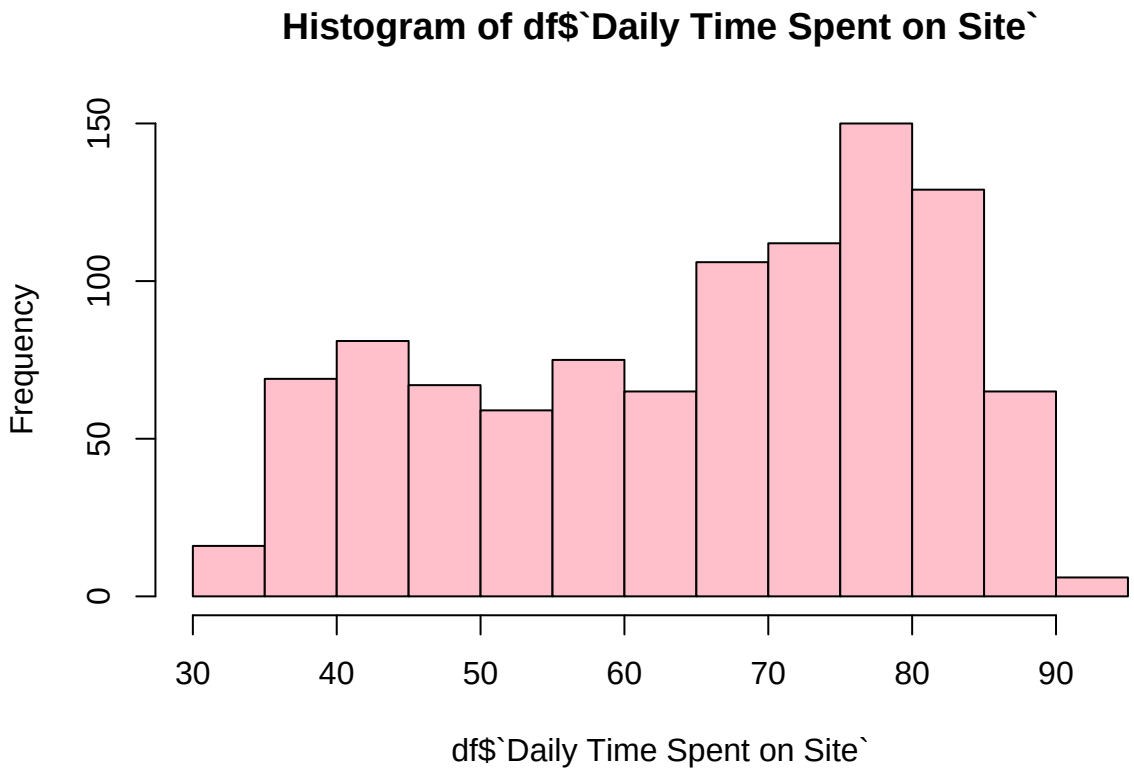
```
median(df$`Daily Time Spent on Site`)
```

```
## [1] 68.215
```

```
#mode
x <- df$`Daily Time Spent on Site`

#sort(x)
names(table(x))[table(x)==max(table(x))]
```

```
## [1] "62.26" "75.55" "77.05" "78.76" "84.53"
```

```
# visualizing the daily time spent on site using histogram
hist(df$`Daily Time Spent on Site`, col = 'pink')
```




```
# majority of the people spent between 75 and 80 hrs on her site.
```

- The users spend an average of 65.002 minutes on the website
- The modal time is “62.26” “75.55” “77.05” “78.76” “84.53”
- The median time is 68.215
- The distribution above is left-skewed

Country

```
# Country where the user belongs
# measures of central tendency
country_x <- df$Country

# levels(country_x) code gives the names of the countries
length(levels(country_x))
```

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

```
#the modal countries in the dataset
names(table(country_x))[table(country_x)==max(table(country_x))]
```

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

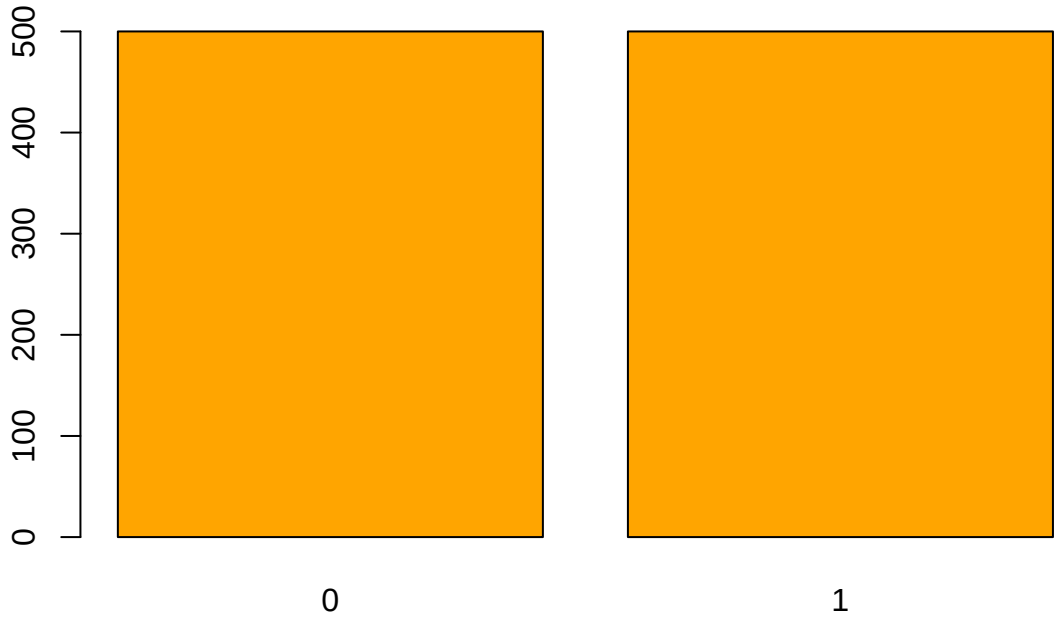
- The most popular countries are: “Czech Republic” “France”

clicked on ads

```
# Visualizing the number of people who clicked on the ads using bar graph
# fetching the clicked on add column
clicked_on_ad <- df$`Clicked on Ad`

#applying the table() function to compute the frequency of clicked on ad column
clicked_on_ad_frequency <- table(clicked_on_ad)
clicked_on_ad_frequency
```

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

```
barplot(clicked_on_ad_frequency, col = 'orange')
```



```
# half the number of people who had an access to her ads clicked on it.
```

BIVARIATE ANALYSIS

```
# Checking for covariance between area income and daily internet usage variables
#assigning the area income column to area_income variable
area_income <- df$`Area Income`

#assigning the daily internet usage column to daily_internet_usage
daily_internet_usage <- df$`Daily Internet Usage`

cov(area_income, daily_internet_usage)
```

```
## [1] 198762.5
```

- There is a positive linear relationship between the two variables

```
# Checking for correlation coefficient between area income and daily internet usage variables
cor(area_income, daily_internet_usage)
```

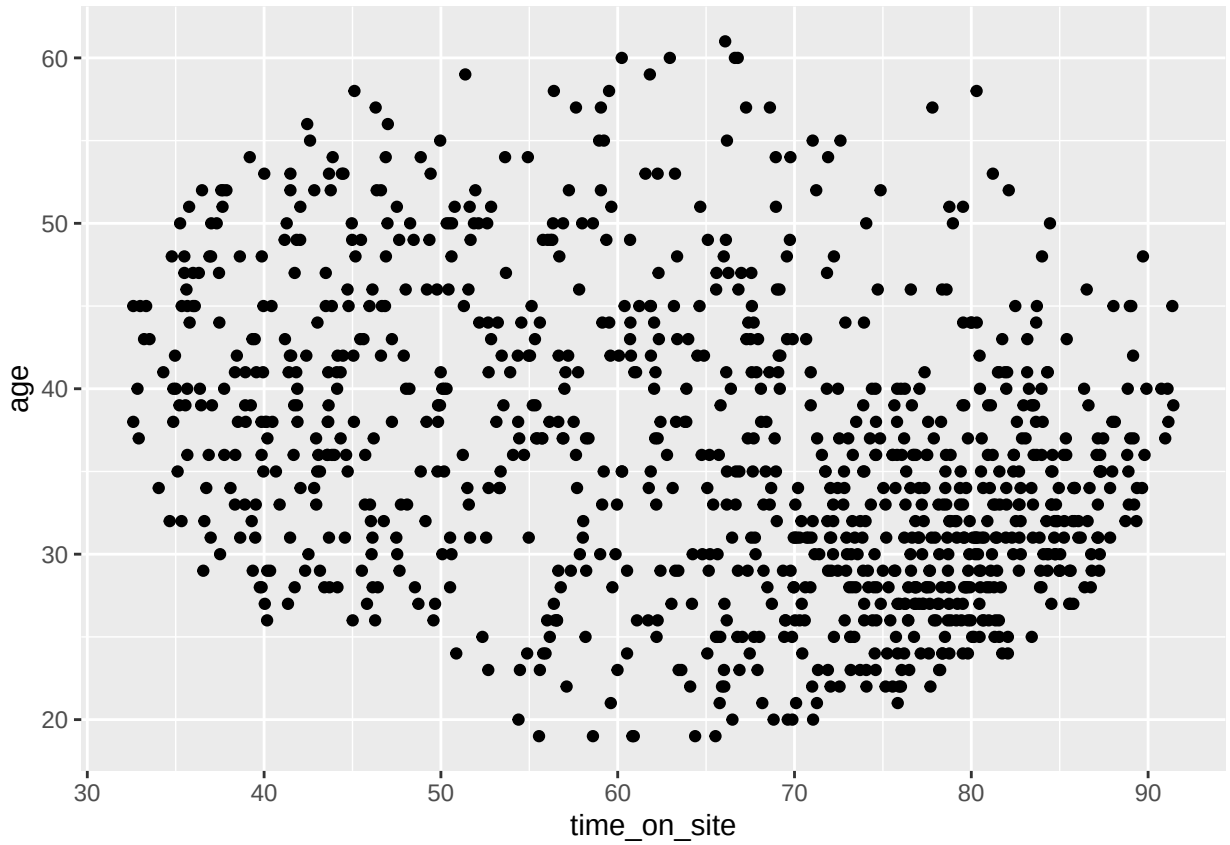
```
## [1] 0.3374955
```

-This indicates that there is a weak linear relationship between the two variables because their correlation coefficient is close to zero

Scatterplots of a few pairs of variables

##Time spent on the site vs age of the user

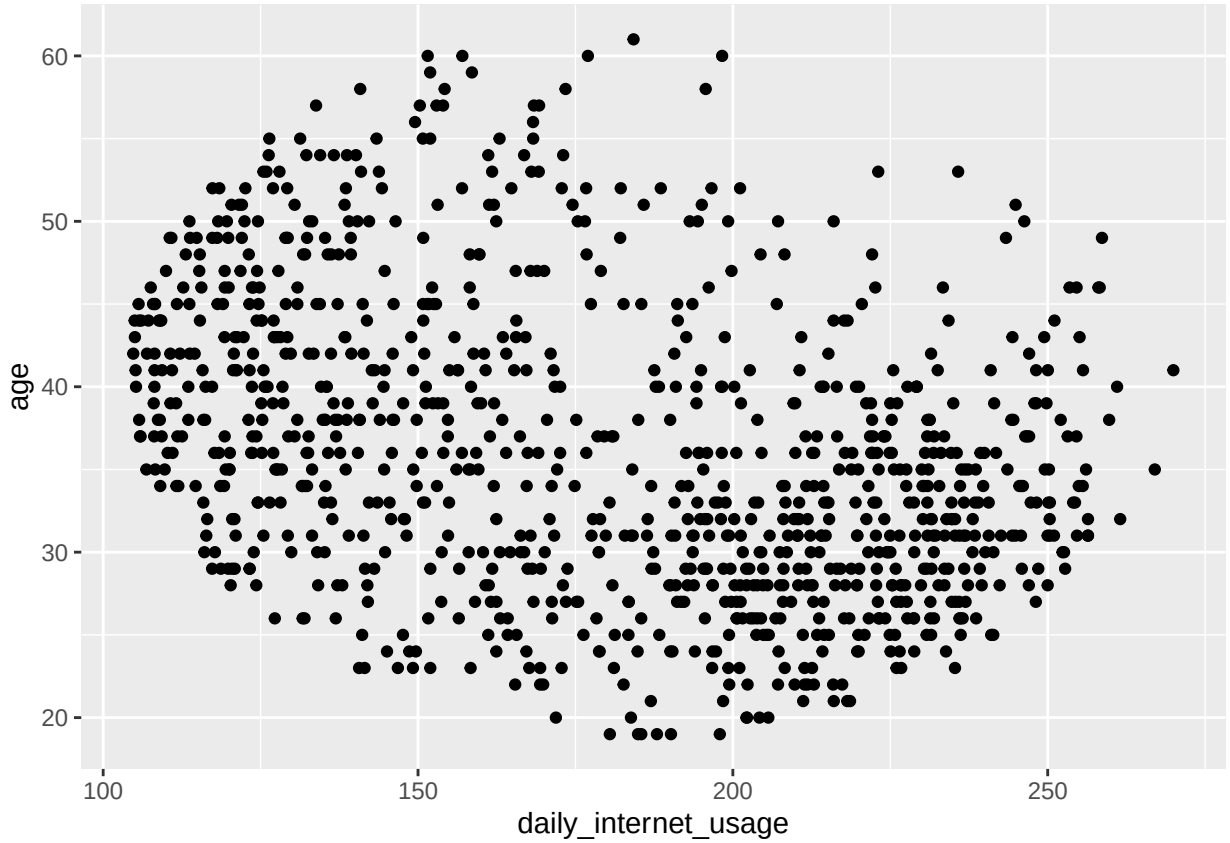
```
#libraries\
library(ggplot2)

#creating the data
time_on_site <- df$`Daily Time Spent on Site`
age <- df$Age
data <- data.frame(time_on_site,age)

#plot
ggplot(data, aes(x=time_on_site, y=age)) + geom_point()
```



- Positive non-linear correlation

Age of the user vs Daily internet usage

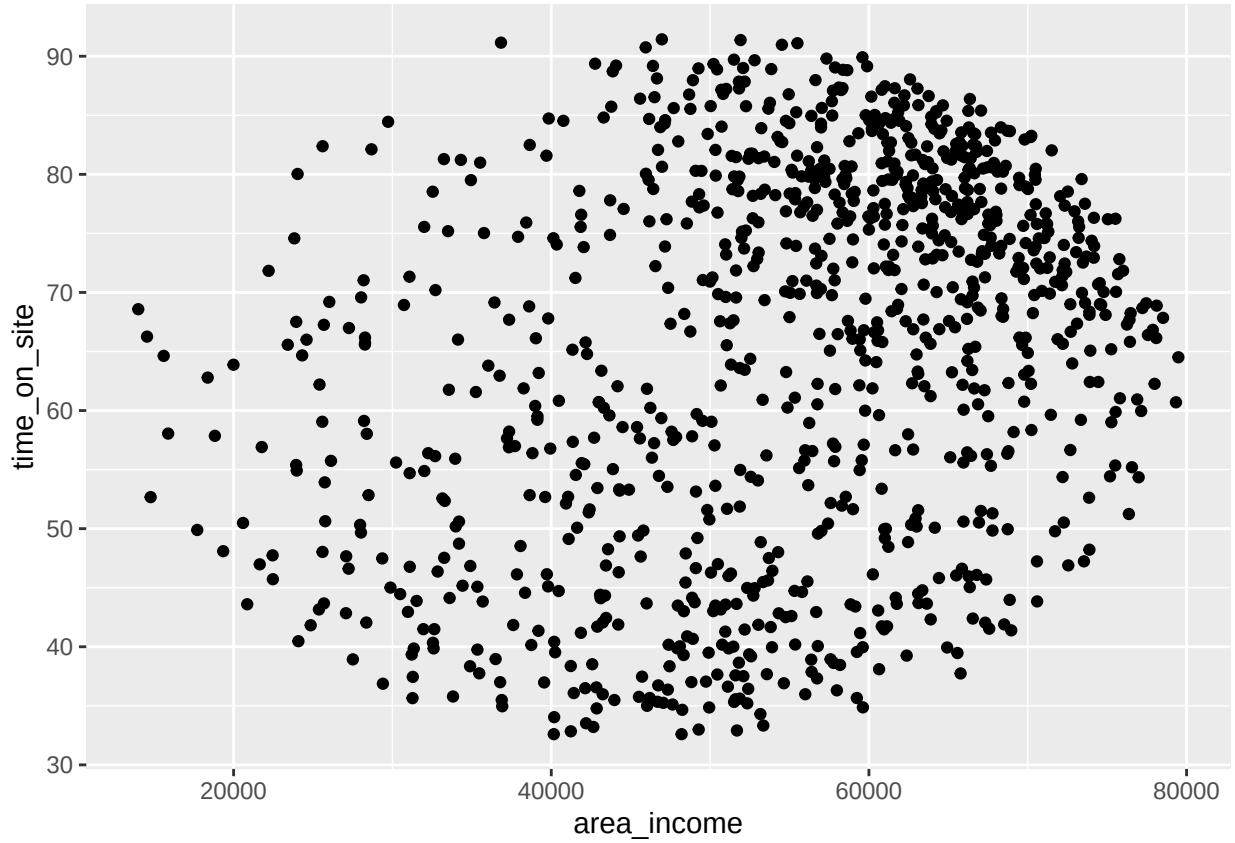
```
data1 <- data.frame(daily_internet_usage,age)

#plot
ggplot(data1, aes(x=daily_internet_usage,y=age)) + geom_point()
```



- The plot shows that there is positive non-linear correlation

Time spent on the site vs Area income

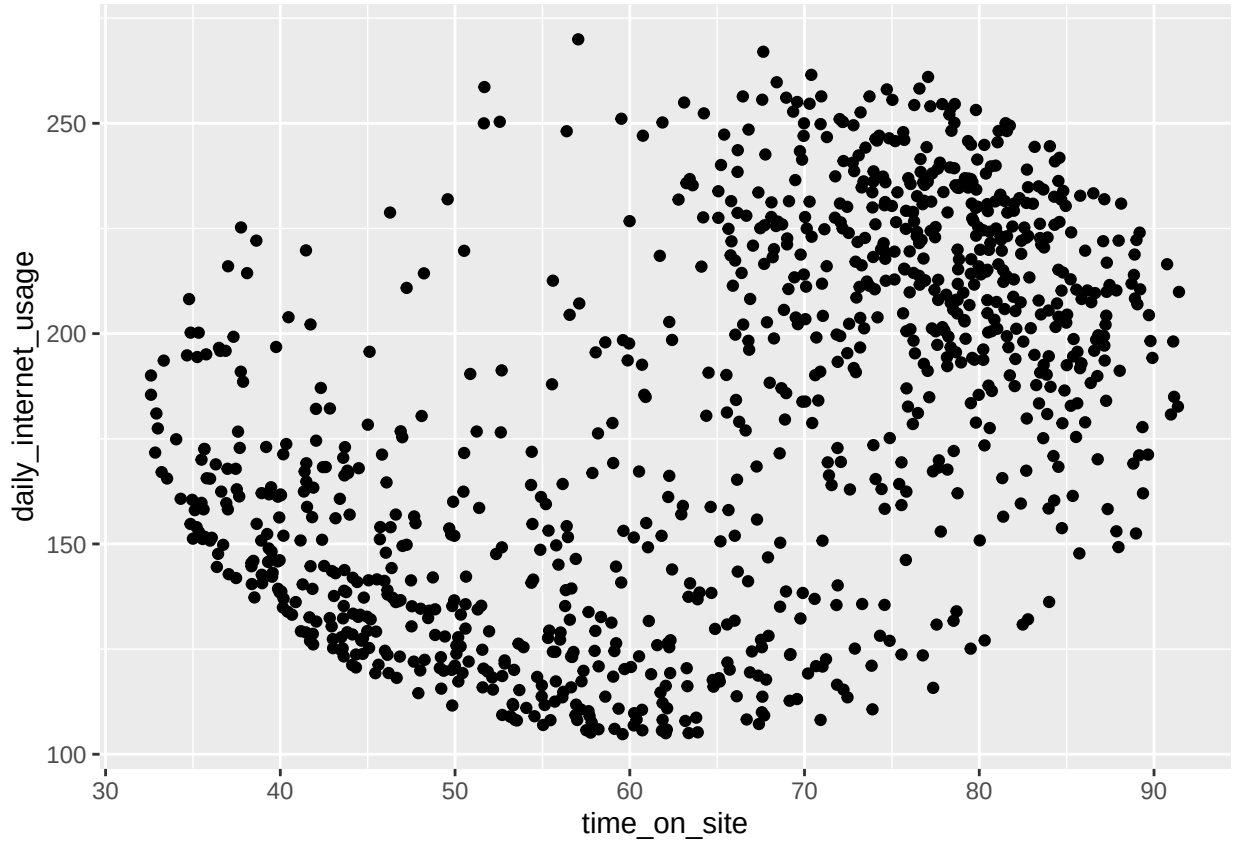
```
data2 <- data.frame(area_income,time_on_site)

#plot
ggplot(data2, aes(x=area_income,y=time_on_site)) + geom_point()
```



- Positive non-linear correlation

Time spent on the site vs Daily internet usage

```
data3 <- data.frame(time_on_site,daily_internet_usage)

#plot
ggplot(data3, aes(x=time_on_site, y=daily_internet_usage)) + geom_point()
```



Gender Vs Clicked on ads

```
library(tidyverse)

# Male respondents who clicked on the ads
dim(df %>% filter(Male == 1, `Clicked on Ad` == 1))
```

```
## [1] 231  10
```

```
# Male respondents who did not on the ads
dim(df %>% filter(Male == 1, `Clicked on Ad` == 0))
```

```
## [1] 250  10
```

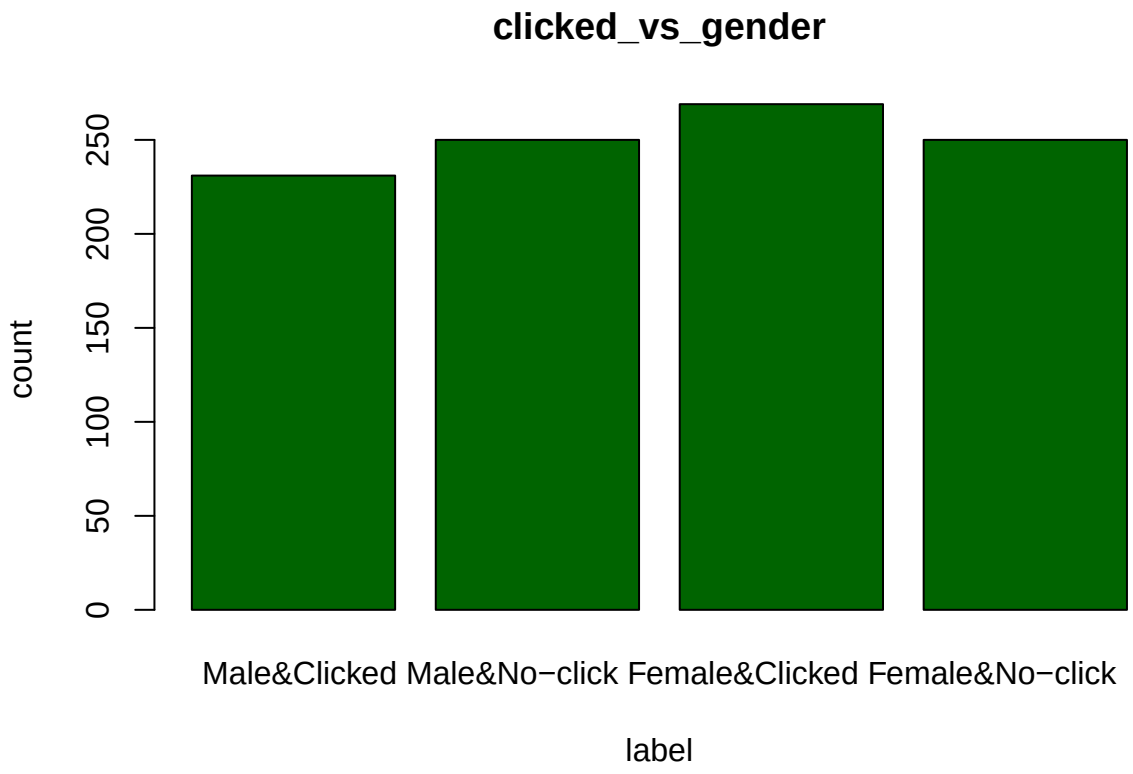
```
# Female respondents who clicked on the ads
dim(df %>% filter(Male == 0, `Clicked on Ad` == 1))
```

```
## [1] 269  10
```

```r
# Female respondents who did not click on the ads
dim(df %>% filter(Male == 0, `Clicked on Ad` == 0))
```

```
## [1] 250  10
```

```r
clicked_vs_gender <- c(231, 250, 269, 250)

# bargraph with added parameters
barplot(clicked_vs_gender, main = 'clicked_vs_gender', xlab = 'label',ylab = 'count', names.arg = c('Ma
```



CONCLUSION

1. The time a user spends on the site does not influence the possibility of clicking on an ad.
2. The gender of respondents who clicked on an ad and those who did not click on an add does not vary much. This means that the Gender of the respondent should be considered in equal measure.
3. Most of the site users who are likely to click on an add earn between 55,000 to 70,000 per month. There are low income earners who click on ads but the majority earn the amount stated above.