online cryptography course project

Dennis Kiarie

8/27/2021

1. Define the Question

1.1 Research Question

Our Research seeks to identify which individuals are most likely to click on ads from the advertisement website.

1.2 Metric of Success

To identify individuals who are relatively more likely to click on course advertisement ads.

1.3 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 Experimental Design

1. Loading Data into RStudio.
2. Checking the Data.
3. Tidying the Data
4. Conducting Exploratory Data Analysis i.e Univariate, Bivariate and Multivariate Analysis.
5. Challenging the Solution by providing insights on how we can make improvements.
6. Recommendations.

1.5 Data Relevance

The data provided was collected in the past from an advertisement website hence its appropriate for our analysis.

The dataset for this Analysis can be found from this link:[http://bit.ly/IPAdvertisingData].

Description of Variables used in this Analysis are :

1.Daily Time Spent on Site; consumer time on site in minutes.

2.Age; customer age in years.

3.Area; geographical area of a consumer.

4.Income; Avg. Income of a consumer.

5.Daily Internet Usage; Avg. minutes a day consumer is on the internet.

6.Ad Topic Line; Headline of the advertisement.

7.City; city of consumer.

8.Male; whether or not a consumer was a male.

9.Country; country of a consumer.

10.Timestamp; Time at which consumer clicked on Ad or closed window

11.Clicked on Ad; 1 indicated clicking on Ad and 0 indicates no clicking on ad.

The dataset used in this Analysis has 1,000 rows and 10 columns.

```
library(tinytex)
```

2.Data Preparation

```
## Importing libraries
#---
#
library(pacman)
pacman :: p_load(pacman, dplyr, GGally, ggplot2, ggthemes, ggvis, httr, lubridate, plotly, rio , rmarkd
theme_set(theme_classic())
options(warn = -1)
```

```
## Loading the data from a csv file
#---
#
ad <- read.csv("C:/Users/Denoo/Downloads/advertising.csv", na.strings = "")

##preview the first 6 rows
#---
#
head(ad)
```

```
##   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
```

```
##preview the last 6 records of the dataset
#---
#
tail(ad)
```

```
##      Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage
## 995                     43.70  28    63126.96               173.01
## 996                     72.97  30    71384.57               208.58
## 997                     51.30  45    67782.17               134.42
## 998                     51.63  51    42415.72               120.37
## 999                     55.55  19    41920.79               187.95
## 1000                    45.01  26    29875.80               178.35
##                            Ad.Topic.Line          City Male
## 995         Front-line bifurcated ability  Nicholasland    0
## 996         Fundamental modular algorithm     Duffystad    1
## 997       Grass-roots cohesive monitoring   New Darlene    1
## 998          Expanded intangible solution South Jessica    1
## 999  Proactive bandwidth-monitored policy   West Steven    0
## 1000      Virtual 5thgeneration emulation   Ronniemouth    0
##                      Country           Timestamp Clicked.on.Ad
## 995                  Mayotte 2016-04-04 03:57:48             1
## 996                  Lebanon 2016-02-11 21:49:00             1
## 997  Bosnia and Herzegovina 2016-04-22 02:07:01             1
## 998                 Mongolia 2016-02-01 17:24:57             1
## 999                Guatemala 2016-03-24 02:35:54             0
## 1000                  Brazil 2016-06-03 21:43:21             1
```

3. Checking the data

```
## we check for the number of rows and columns
#---
#
cat("Rows:", nrow(ad), "\nCols:", ncol(ad))
```

```
## Rows: 1000
## Cols: 10
```

```
##we check the type
#---
#
class(ad)
```

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

```
##we check if datatypes are appropriate
#---
#
glimpse(ad)
```

```
## Rows: 1,000
## Columns: 10
```

```
## $ Daily.Time.Spent.on.Site <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, 88.~
## $ Age                      <int> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49, 3~
## $ Area.Income              <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 73889~
## $ Daily.Internet.Usage     <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 226.7~
## $ Ad.Topic.Line            <chr> "Cloned 5thgeneration orchestration", "Monito~
## $ City                     <chr> "Wrightburgh", "West Jodi", "Davidton", "West~
## $ Male                     <int> 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, ~
## $ Country                  <chr> "Tunisia", "Nauru", "San Marino", "Italy", "I~
## $ Timestamp                <chr> "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, 1, 0, 1, 0, 1, ~
```

The column for Timestamp has an inappropriate datatype which will be rectified during data cleaning process.

```
##we check the column names for easier reference
#---
#
names(ad)
```

```
##  [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"
```

```
## we Check for unique characters
sapply(ad, function(x) length(unique(x)))
```

```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##                      900                       43                     1000
##     Daily.Internet.Usage           Ad.Topic.Line                     City
##                      966                     1000                      969
##                     Male                  Country                Timestamp
##                        2                      237                     1000
##            Clicked.on.Ad
##                        2
```

4. Tidying the data

```
##we change the column names to lowercase for easier manipulation
#---
#
colnames(ad) = tolower(colnames(ad))
colnames(ad)
```

```
##  [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"
```

```
##we replace spaces in column names for easier manipulation
#---

names(ad) = str_replace_all(names(ad), c(' ' = '_'))
names(ad)
```

```
##  [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"
```

```
##we rename column male to gender
#---
#

colnames(ad)[colnames(ad) == 'male'] = 'gender'
```

```
## we check for the datatypes for each column
#---
#
columns = names(ad)
for (column in seq(length(names(ad)))){
    print(columns[column])
    print(class(ad[, column]))
    cat('\n')
}
```

```
## [1] "daily.time.spent.on.site"
## [1] "numeric"
##
## [1] "age"
## [1] "integer"
##
## [1] "area.income"
## [1] "numeric"
##
## [1] "daily.internet.usage"
## [1] "numeric"
##
## [1] "ad.topic.line"
## [1] "character"
##
## [1] "city"
## [1] "character"
##
## [1] "gender"
## [1] "integer"
##
## [1] "country"
## [1] "character"
##
```

```
## [1] "timestamp"
## [1] "character"
##
## [1] "clicked.on.ad"
## [1] "integer"
```

```
## we create a list of categorical columns
#---
#
cat_cols = c("ad_topic_line", "city", "gender", "country", "clicked_on_ad" )

## we create a List of numerical columns
num_cols = c("daily_time_spent_on_site", "age", "area_income", "daily_internet_usage")
```

```
## we Change the datatypes
#---
#
ad$gender <- as.factor(ad$gender)
ad$clicked.on.ad <- as.factor(ad$clicked.on.ad)

##we check for datatypes
sapply(ad, class)
```

```
## daily.time.spent.on.site                      age              area.income
##                "numeric"                "integer"                "numeric"
##     daily.internet.usage            ad.topic.line                     city
##                "numeric"              "character"              "character"
##                   gender                  country                timestamp
##                 "factor"              "character"              "character"
##            clicked.on.ad
##                 "factor"
```

```
##we check for duplicates
#---
#
anyDuplicated(ad)
```

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

```
#There are no duplicates
```

```
##we check for missing values
#---
#
colSums(is.na(ad))
```

```
## daily.time.spent.on.site                      age              area.income
##                        0                        0                        0
##     daily.internet.usage            ad.topic.line                     city
##                        0                        0                        0
##                   gender                  country                timestamp
```

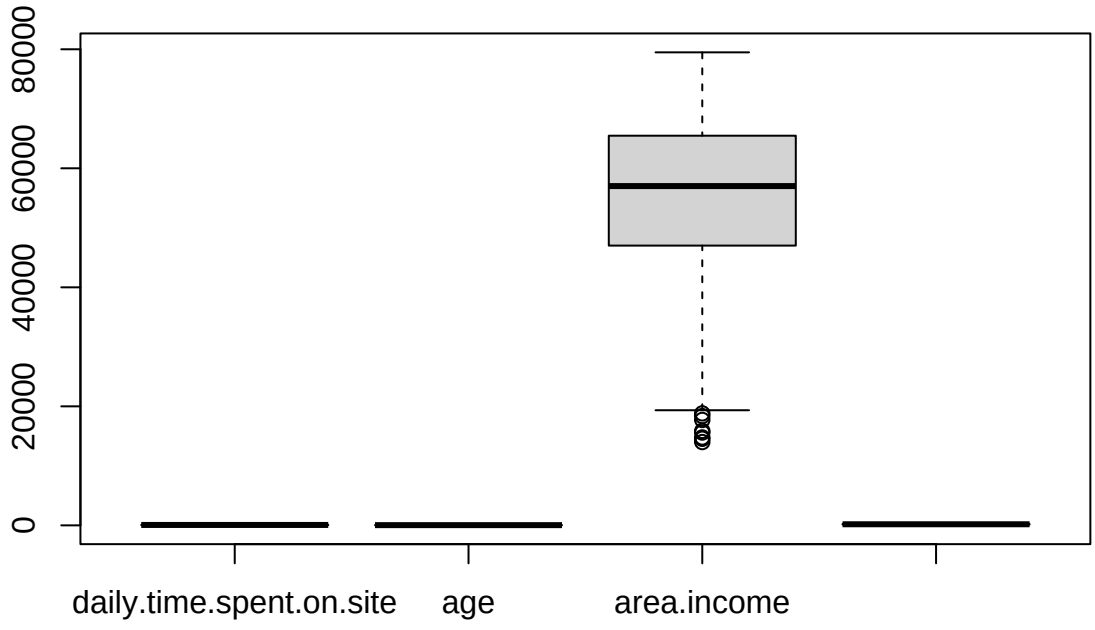
```
##                         0                         0                         0
##            clicked.on.ad
##                         0
```

```
#we have no missing values
```

```
##we check for outliers in numerical columns
#---
#
df <- subset(ad, select = -c(ad.topic.line,city,    gender, country,    timestamp,  clicked.on.ad))

##preview the data
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
```

```
## we Plot boxplots to check for outliers
#---
#
boxplot(df)
```



```
# Only area.income variable has a few outliers of individuals  earning less than 20000.
#We investigate them so as to see whether they are legit or not.
```

```
## we select all the rows with outliers
#---
#
outliers <- df %>% dplyr::filter(area.income < 20000)
outliers
```

```
##    daily.time.spent.on.site age area.income daily.internet.usage
## 1                     49.89  39    17709.98               160.03
## 2                     63.88  38    19991.72               136.85
## 3                     48.09  33    19345.36               180.42
## 4                     57.86  30    18819.34               166.86
## 5                     64.63  45    15598.29               158.80
## 6                     58.05  32    15879.10               195.54
## 7                     66.26  47    14548.06               179.04
## 8                     68.58  41    13996.50               171.54
## 9                     52.67  44    14775.50               191.26
## 10                    62.79  36    18368.57               231.87
```

Conclusion

We choose to work with outliers.This because individuals earn different amount of income. Also, from our dataset the currency of income is not defined and we are not told if they are annual,monthly or weekly income. Therefore, removing this outliers will affect the accuarcy of the data analysis and the result will be inconclusive.

```
## we Change column timestamp to datetime
#---
#
ad$Timestamp <- as.POSIXct(ad$timestamp, "%Y-%m-%d %H:%M:%S",tz = "GMT")

##preview the data
head(ad)
```

```
##   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 gender    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           Timestamp
## 1 2016-03-27 00:53:11             0 2016-03-27 00:53:11
## 2 2016-04-04 01:39:02             0 2016-04-04 01:39:02
```

```
## 3 2016-03-13 20:35:42             0 2016-03-13 20:35:42
## 4 2016-01-10 02:31:19             0 2016-01-10 02:31:19
## 5 2016-06-03 03:36:18             0 2016-06-03 03:36:18
## 6 2016-05-19 14:30:17             0 2016-05-19 14:30:17
```

```
## we split time and date from Timestamp
#---
#
ad$date = format(ad$Timestamp, "%y/%m/%d")
ad$time = format(ad$Timestamp, "%H:%M:%S")
ad$date <- as.Date(ad$date)

#preview the data
head(ad)
```

```
##   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 gender    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           Timestamp       date     time
## 1 2016-03-27 00:53:11             0 2016-03-27 00:53:11 0016-03-27 00:53:11
## 2 2016-04-04 01:39:02             0 2016-04-04 01:39:02 0016-04-04 01:39:02
## 3 2016-03-13 20:35:42             0 2016-03-13 20:35:42 0016-03-13 20:35:42
## 4 2016-01-10 02:31:19             0 2016-01-10 02:31:19 0016-01-10 02:31:19
## 5 2016-06-03 03:36:18             0 2016-06-03 03:36:18 0016-06-03 03:36:18
## 6 2016-05-19 14:30:17             0 2016-05-19 14:30:17 0016-05-19 14:30:17
```

```
##we drop the column Timestamp
#---
#
final_df = subset(ad, select = -c(Timestamp))
#preview the data
head(final_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 gender    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       date     time
## 1 2016-03-27 00:53:11             0 0016-03-27 00:53:11
## 2 2016-04-04 01:39:02             0 0016-04-04 01:39:02
## 3 2016-03-13 20:35:42             0 0016-03-13 20:35:42
## 4 2016-01-10 02:31:19             0 0016-01-10 02:31:19
## 5 2016-06-03 03:36:18             0 0016-06-03 03:36:18
## 6 2016-05-19 14:30:17             0 0016-05-19 14:30:17
```

```
## we use frequency tables to check data distribution in gender column
#---
#
table(final_df$gender)
```

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

Conclusion

0 represent female and 1 represents male. The data we used for analysis had a slight higher number of females than males hence gender variable is fairly distributed.

```
##we rename the target variable
#---
#
levels(final_df$clicked.on.ad) = c("Yes", "No")

#preview the data
head(final_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 gender    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       date     time
## 1 2016-03-27 00:53:11           Yes 0016-03-27 00:53:11
## 2 2016-04-04 01:39:02           Yes 0016-04-04 01:39:02
## 3 2016-03-13 20:35:42           Yes 0016-03-13 20:35:42
```

```
## 4 2016-01-10 02:31:19           Yes 0016-01-10 02:31:19
## 5 2016-06-03 03:36:18           Yes 0016-06-03 03:36:18
## 6 2016-05-19 14:30:17           Yes 0016-05-19 14:30:17
```

5. Exploratory Data Analysis

5.1 Univariate Analysis

We decided to analyze all columns individually so as to pay close attention to the findings

```
time <- final_df$`daily.time.spent.on.site`
age <- final_df$age
income <- final_df$`area.income`
usage <- final_df$`daily.internet.usage`
```

5.1.1 Age

Measure of Central Tendency

```
summary(time)
```

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

```
## we calculate the mode
#---
#
getmode <- function(v) #getmode is the function name
  {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

#we Calculate the mode using the user function.
mode <- getmode(time)
print(mode)
```

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

Measure of Dispersion

```
##install package moments
#---
#
library(moments)
```

```
##The Variance
#---
#
var(time)
```

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

```r
##The standard deviation
#---
#
sd(time)
```

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

```r
##The range of the Variable
#---
#
range(time)
```

```
## [1] 32.60 91.43
```

```r
##The Interquartile Range
#---
#
IQR(age)
```

```
## [1] 13
```

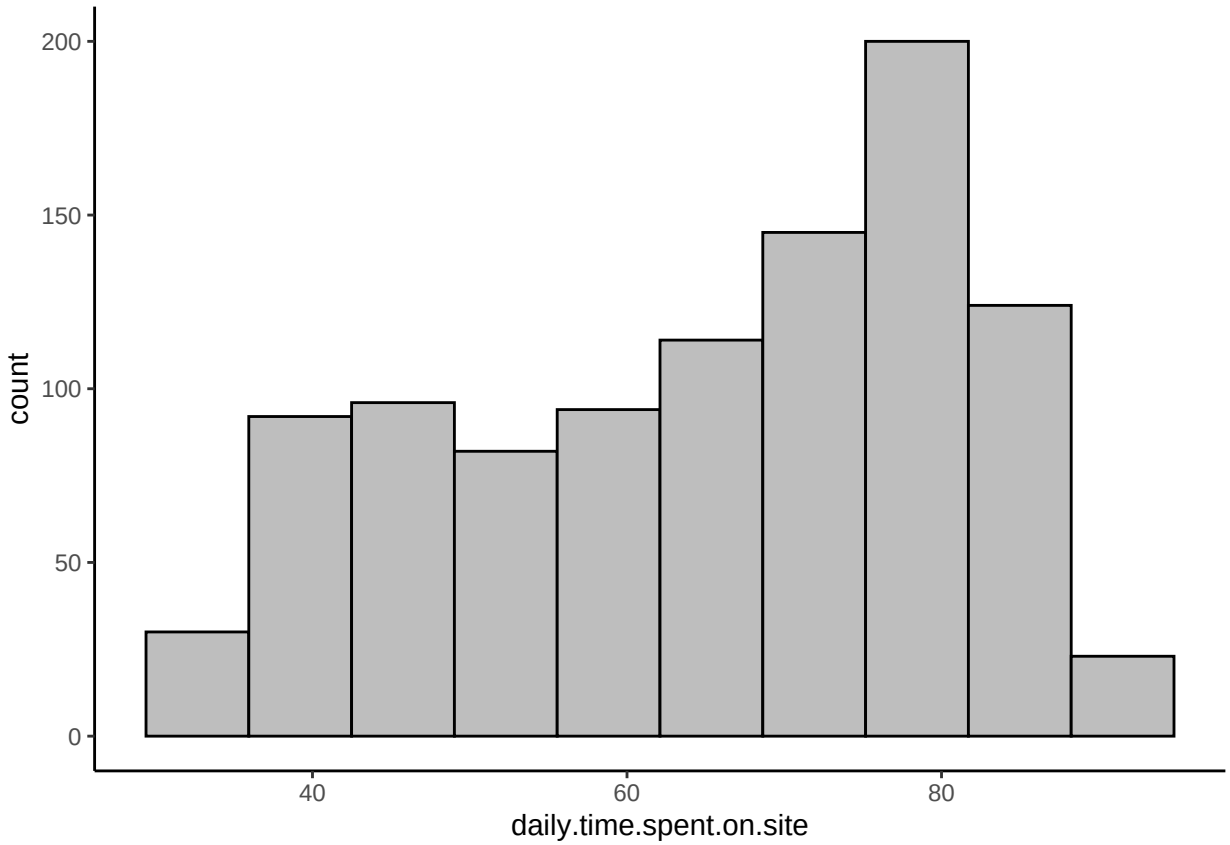
```r
##The Skew of the column
skewness(time)
```

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

```r
##The Kurtosis
#---
#
kurtosis(time)
```

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

```r
##Histogram with density plot
#---
#
ggplot(final_df, aes(x=`daily.time.spent.on.site`)) +
 geom_histogram(colour="black", fill="grey",bins=10)
```



Conclusion

From the variable time we observed he following:

1.The mean amount of time that users spent on the site was 65 minutes while the maximum amount of time a user spent on the site was 91.43 minutes.

2.The Variance of the column was 251.3371 with a standard deviation of 15.85361.

3.The data was negatively but fairly symmetrical with a value of -0.3712026 and the distribution can be categorized as platykurtic with a kurtosis value of 1.903942.

4.From the Histogram we deduced that a pproxiamtely 125 users spent over 80 minutes daily on the site , and many users spent over 60 minutes on the site daily.

5.1.2 Age

Measure of central Tendency

```
summary(age)
```

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

```
##we calculate the mode
#---
#
getmode <- function(v) #getmode is the function name
  {
```

```
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
## we Calculate the mode using the user function.
mode <- getmode(age)
print(mode)
```

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

Measure of Disperson

```
##The Variance
#---
#
var(age)
```

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

```
##The standard deviation
#---
#
sd(age)
```

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

```
##The range of the Variable
#---
#
range(age)
```

```
## [1] 19 61
```

```
##The Interquartile Range
#---
#
IQR(age)
```

```
## [1] 13
```

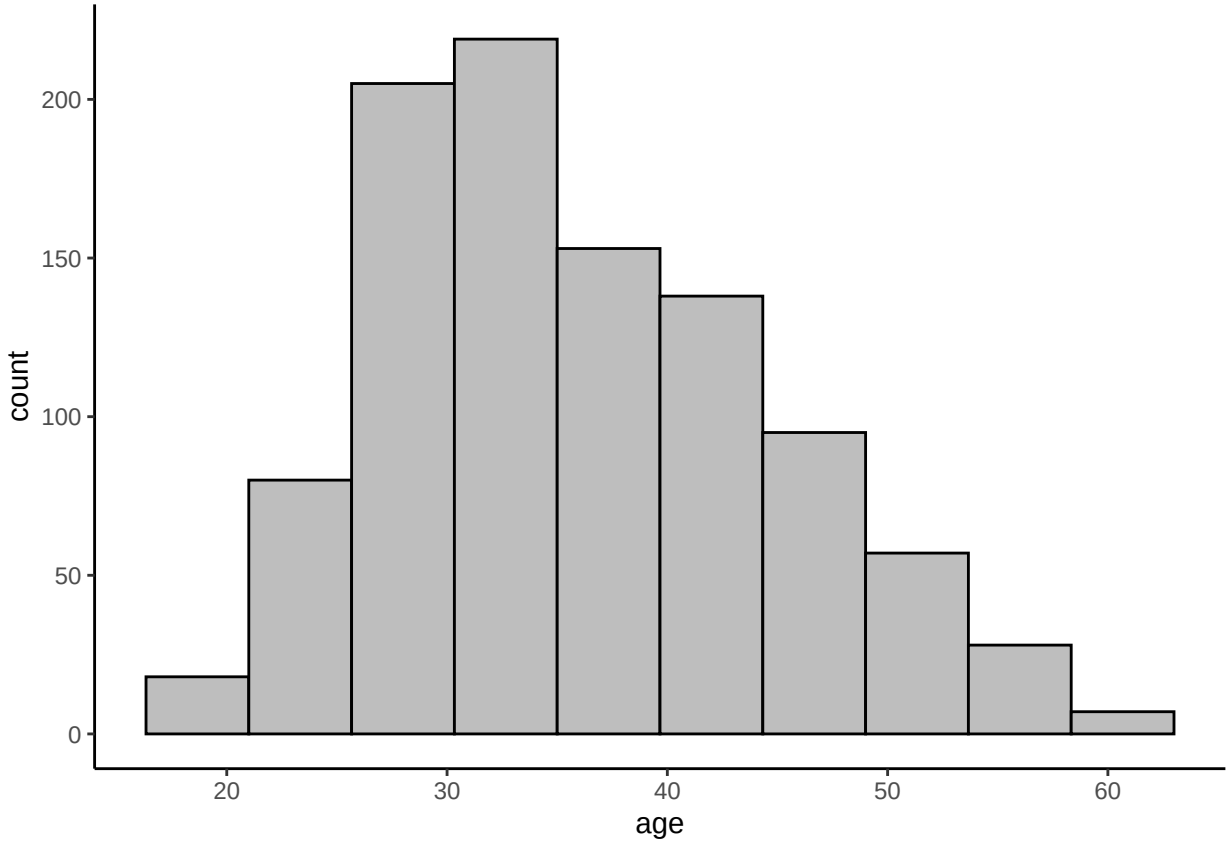
```
##The Skew of the column
skewness(age)
```

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

```
##The Kurtosis
#---
#
kurtosis(age)
```

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

```
##Histogram with density plot
#---
#
ggplot(final_df, aes(x=`age`)) +
 geom_histogram(colour="black", fill="grey",bins=10)
```



Conclusion

1.The mean age of the consumers was 36.01 while the median age was 35 and the modal age was 31.

2.The variable was positively skewed and fairly symmetrical with a skew value of 0.4777052 , the distribution was platykurtic with a kurtosis value of 2.595482

3.The age of the consumers ranged between 19 and 61 , with majority of the users ranging between 30 and 45.

4.The interquartile age for the upper and lower quartile was 13.

5.1.3 Area Income

Measure of central Tendency

```
summary(income)
```

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

```r
##we calculate the mode
#---
#
getmode <- function(v) #getmode is the function name
  {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
## we Calculate the mode using the user function.
mode <- getmode(income)
print(mode)
```

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

Measure of Dispersion

```r
##The Variance
#---
#
var(income)
```

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

```r
##we check the standard deviation
#---
#
sd(income)
```

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

```r
##we check the range
#---
#
range(income)
```

```
## [1] 13996.5 79484.8
```

```r
##we check Interquartile Range
#---
#
IQR(income)
```

```
## [1] 18438.83
```

```r
##we check the skewness
#---
3
```

```
## [1] 3
```

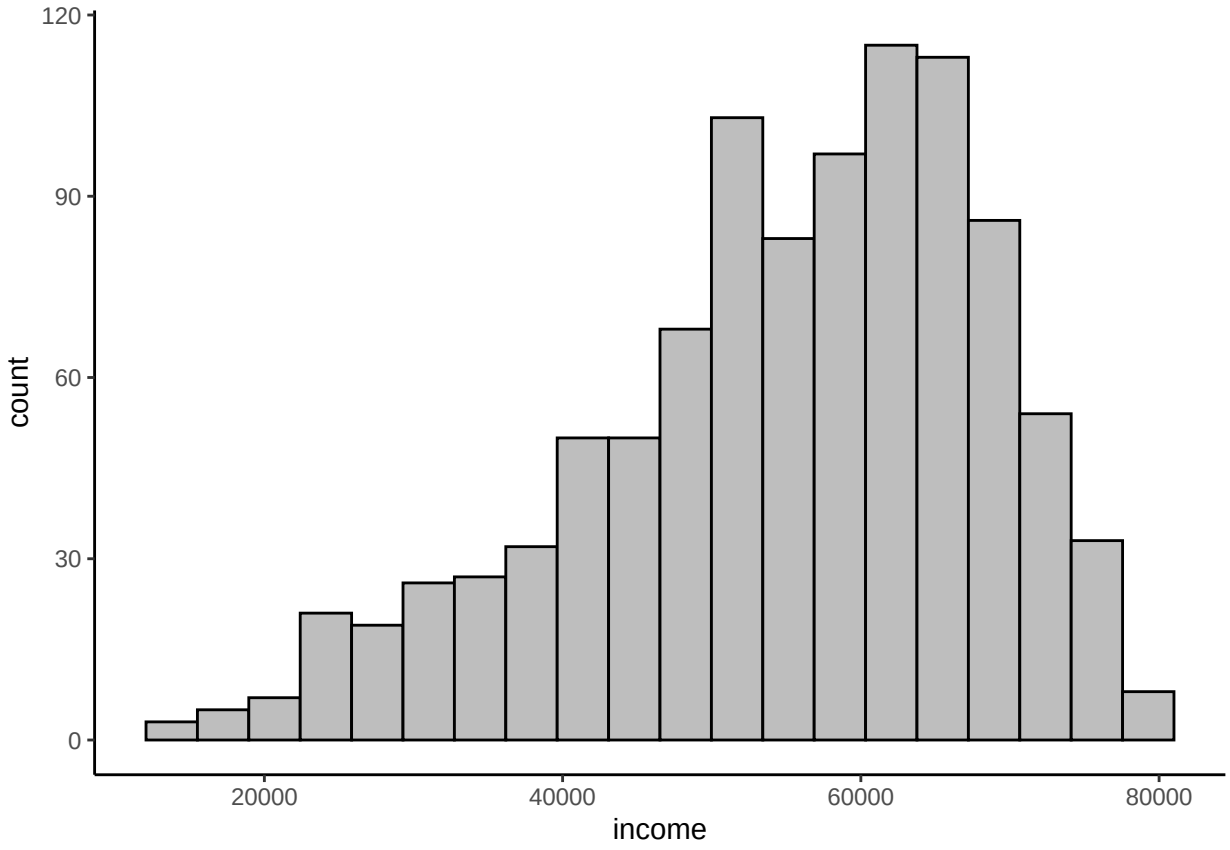
```r
skewness(income)
```

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

```r
##we check for kurtosis
#---
#
kurtosis(income)
```

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

```r
##Histogram with density plot
#---
#
ggplot(final_df, aes(x=`income`)) +
 geom_histogram(colour="black", fill="grey",bins=20)
```



Conclusion

1.The mean income of the users was 55,000 while the median income was 57,012 and the modal income was 61,833.9.

2.The income of the site users ranged from 13,996 to 79,485 , with the interquarile range being 18483.83.

3.The data had a negative skew and was moderately skewed with a value of -0.6484229, the data had a platykurtic distribution with a value of 2.894694.

5.1.4 Daily Internet Usage

Measure of central Tendency

```
summary(usage)
```

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

```
##we calculate the mode
#---
#
getmode <- function(v) #getmode is the function name
  {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
## we Calculate the mode using the user function.
mode <- getmode(usage)
print(mode)
```

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

Measure of Disperson

```
##we check for standard deviation
#---
#
sd(usage)
```

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

```
##we check for variance
#---
#
var(usage)
```

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

```
##we check for range
#---
#
range(usage)
```

```
## [1] 104.78 269.96
```

```
##we check for Interquatile range
#---
#
IQR(usage)
```

```
## [1] 79.9625
```

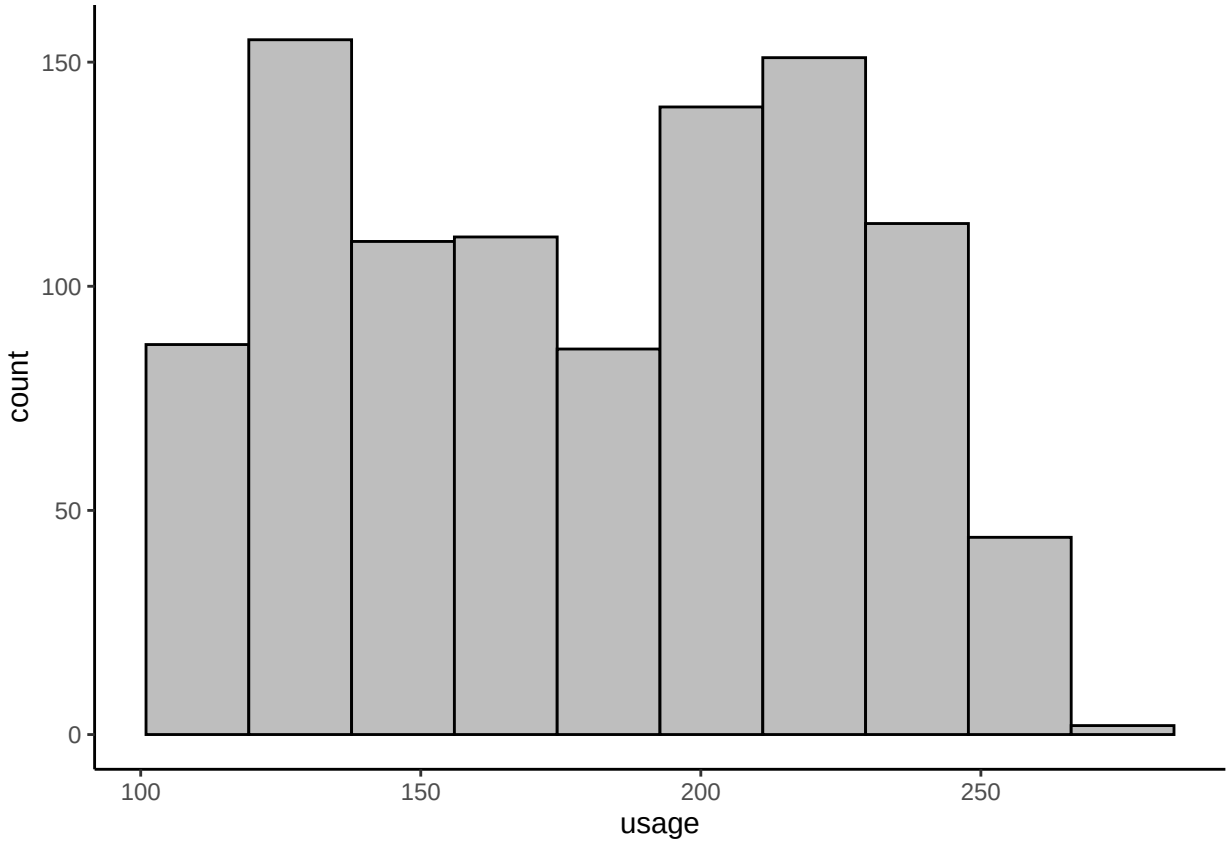
```
##we check the skewness
#---
#
skewness(usage)
```

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

```
##we check for the kurtosis
#---
#
kurtosis(usage)
```

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

```
##Histogram with density plot
#---
#
ggplot(final_df, aes(x=`usage`)) +
 geom_histogram(colour="black", fill="grey",bins=10)
```



Conclusion

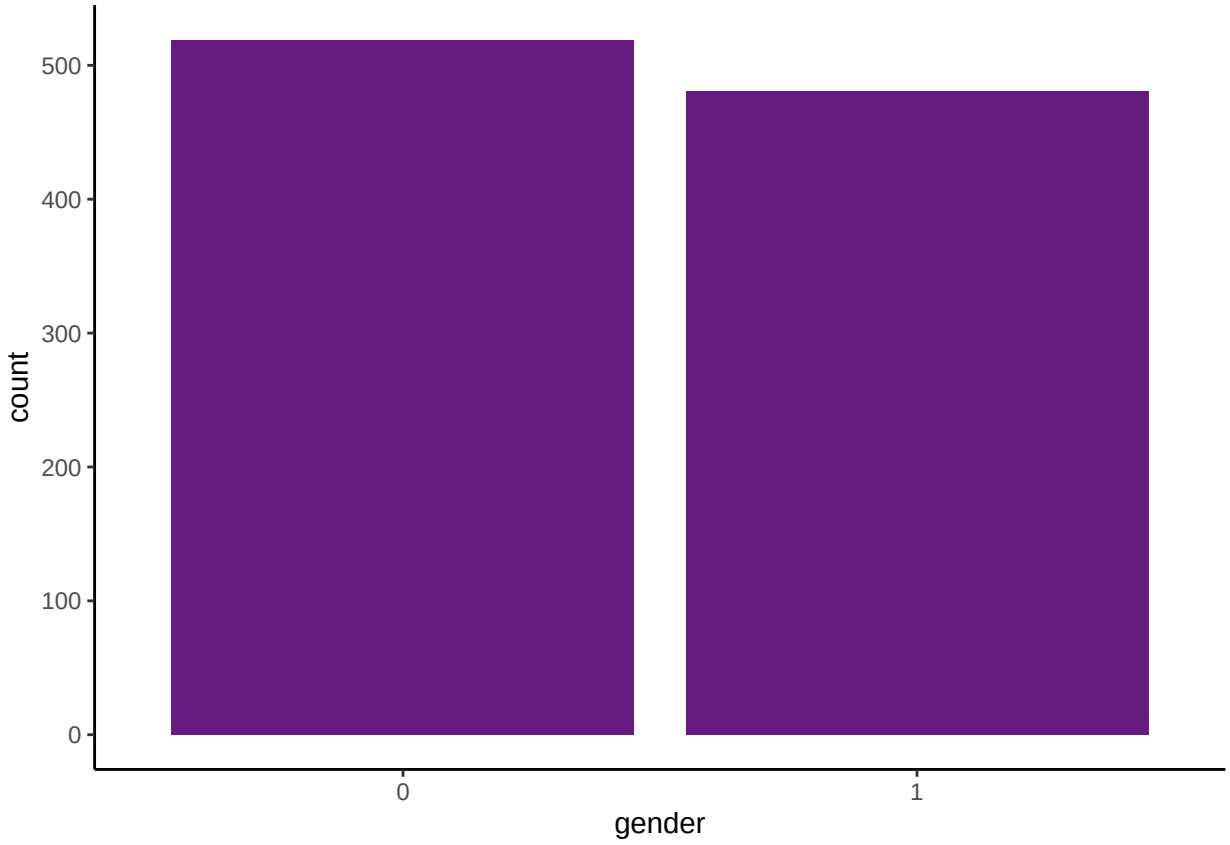
1.The Average Hours spent by users on the Internet is 180 minutes while the median is 183.1 and the mode is 167.22.

2.The Interquartile Range is 79.9625.

3.The data is negatively but fairly skewed with a skew value of -0.03343681 and the data is platykurtic with a value of 1.727701.

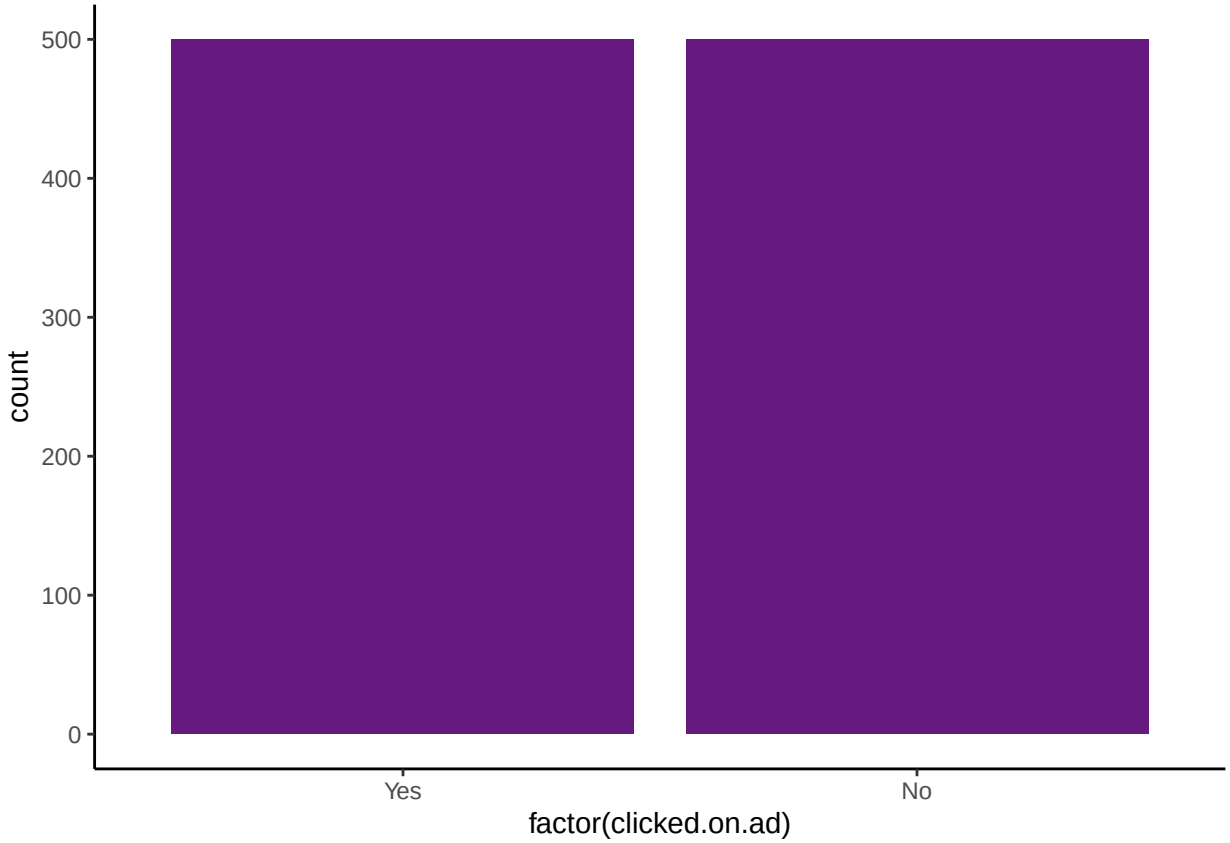
Countplots for the Categorical variables_

```
##we plot the countplot for the variable gender
#---
#
ggplot(final_df, aes(x=gender)) + geom_bar(fill=rgb(0.4,0.1,0.5))
```




```
# 0 represents female and 1 male
#From the plotted countplots the number of female was slightly higher than that of male.
```

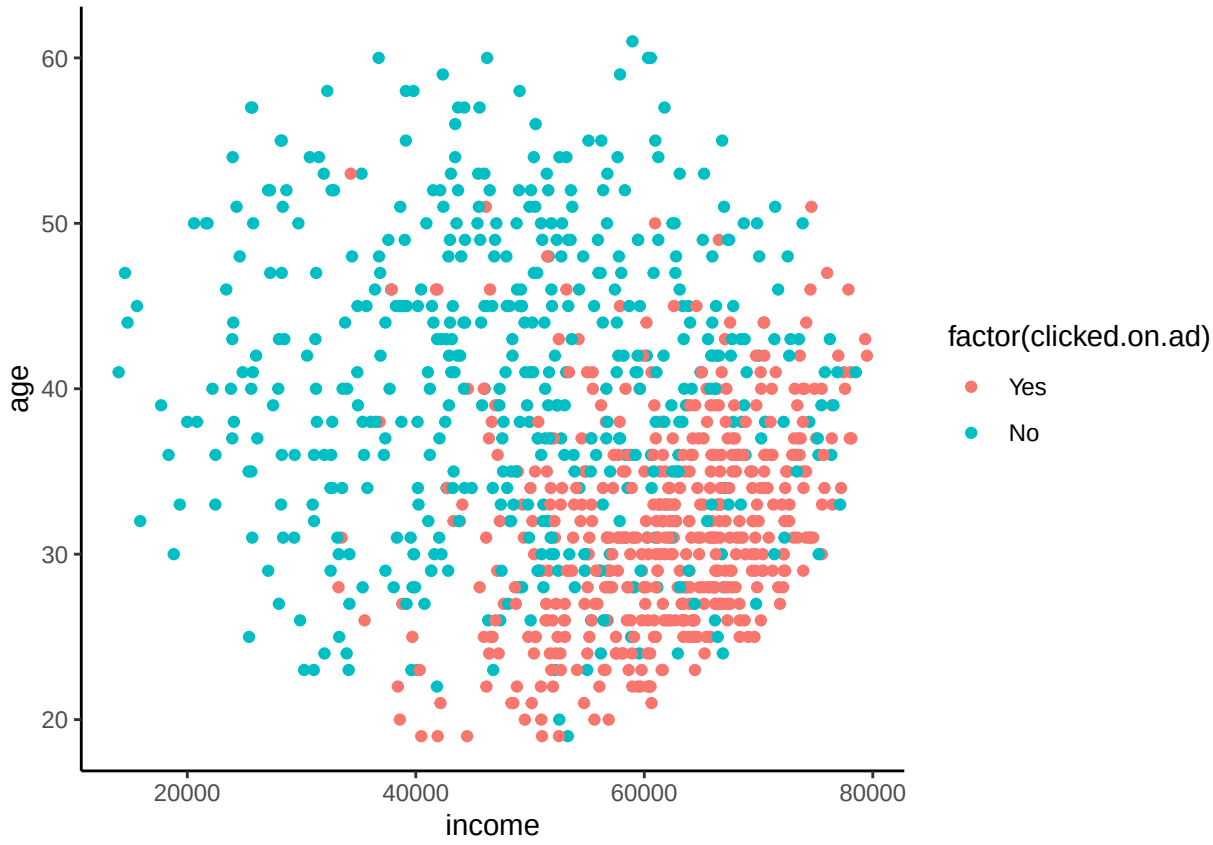
```
##we plot the countplot for the target variable i.e clicked.on.ad
#---
#
ggplot(final_df, aes(x=factor(`clicked.on.ad`))) + geom_bar( fill=rgb(0.4,0.1,0.5))
```



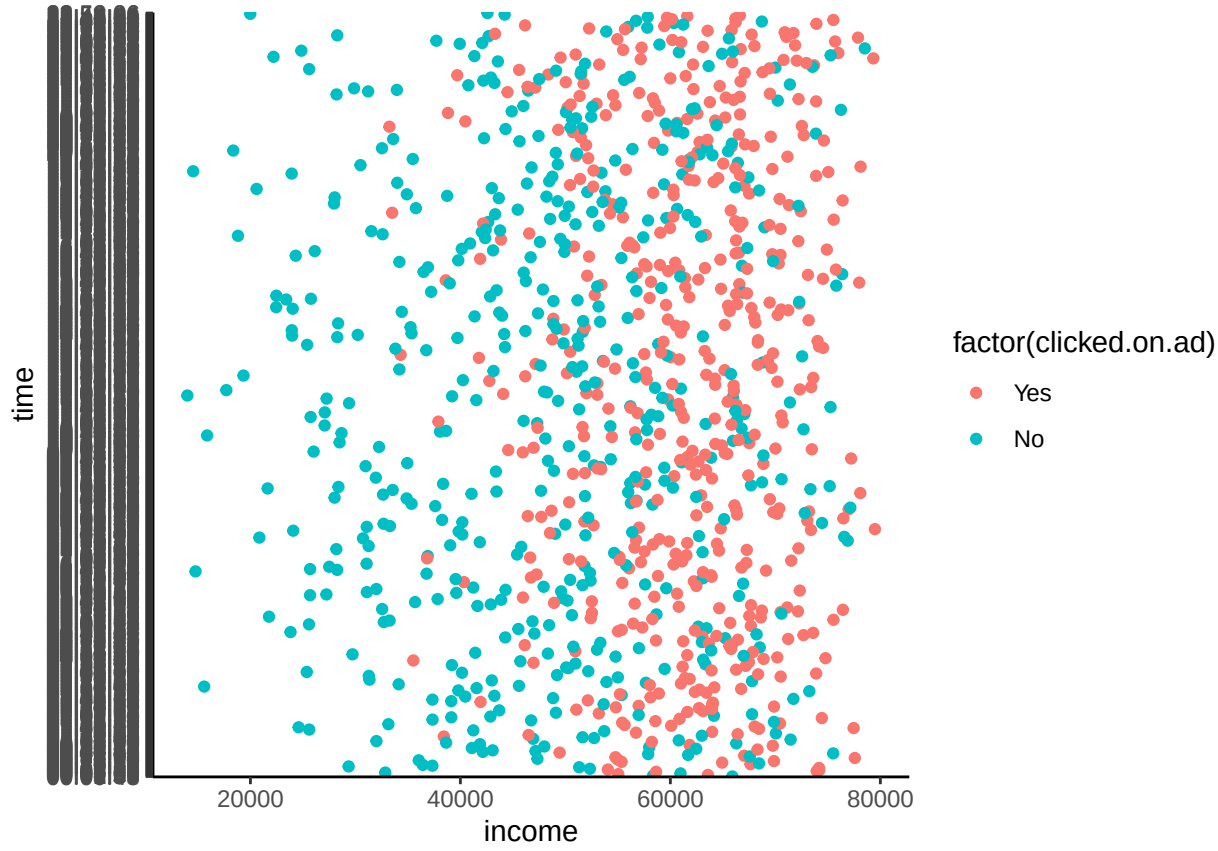

```
#We observed that the number of users on the site who clicked on the ad is equal to those that did #not
```

5.2 Bivariate Analysis

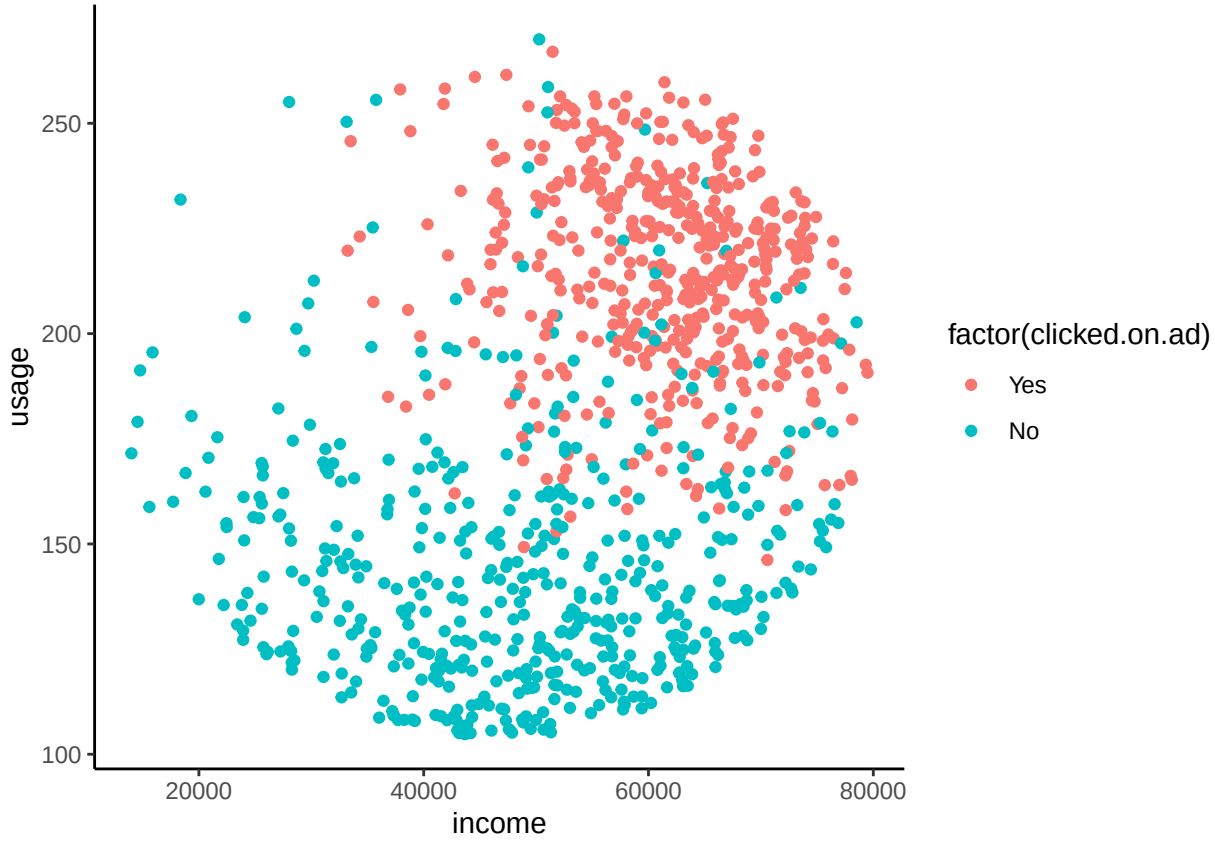
```
##Scatter plot for income and age
#---
#

ggplot(final_df, aes(x=income , y = age )) + geom_point(aes(colour= factor(`clicked.on.ad`)))
```



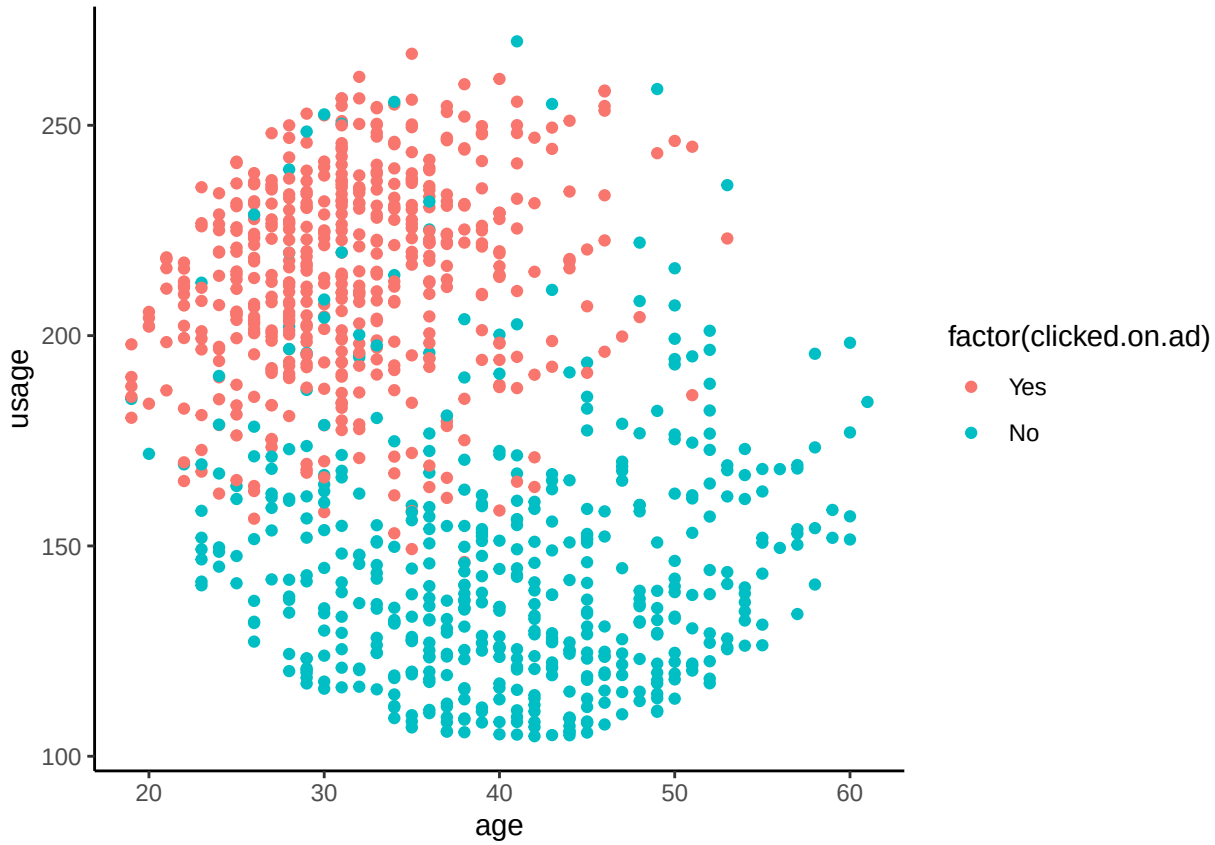

```
##Scatter plot for income and time
#---
#
ggplot(final_df, aes(x=income , y = time )) + geom_point(aes(colour= factor(`clicked.on.ad`)))
```



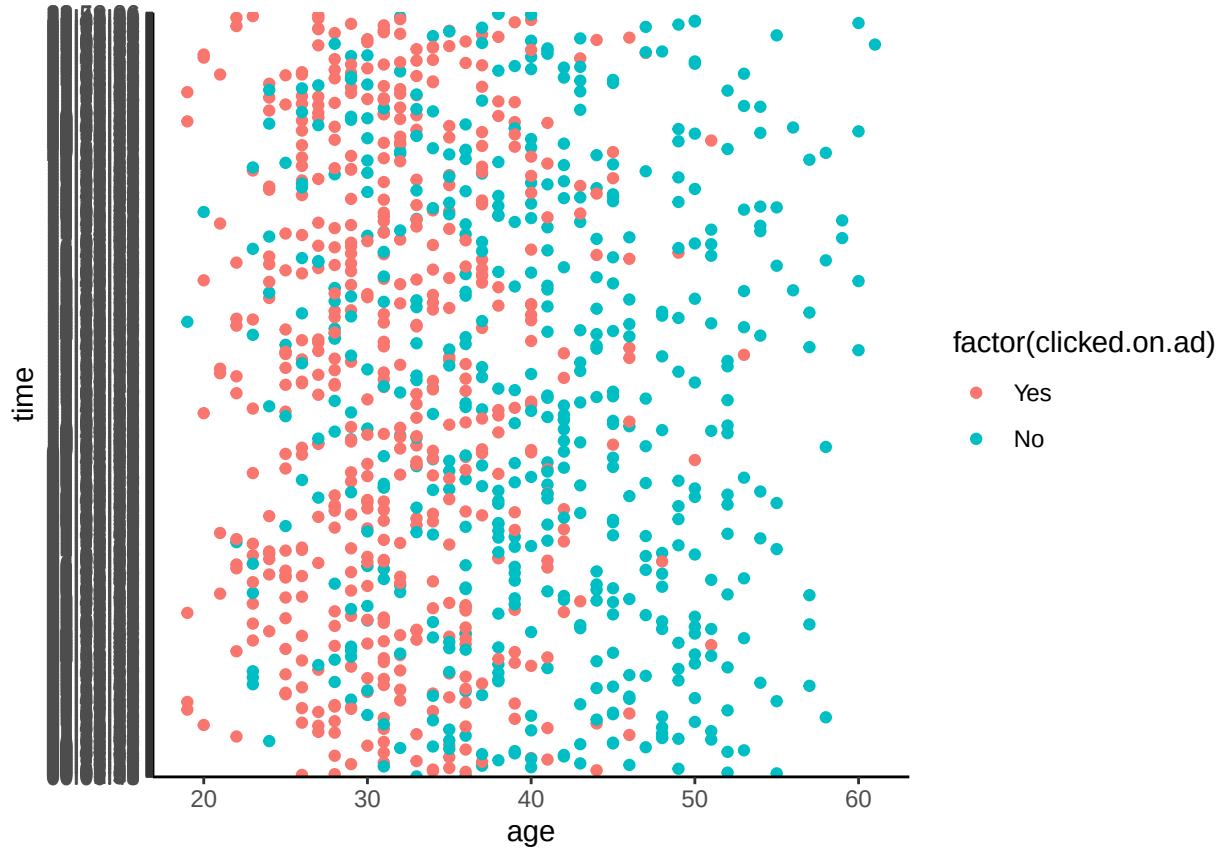

```
##Scatter plot for income and usage
#---
#
ggplot(final_df, aes(x=income , y = usage )) + geom_point(aes(colour= factor(`clicked.on.ad`)))
```



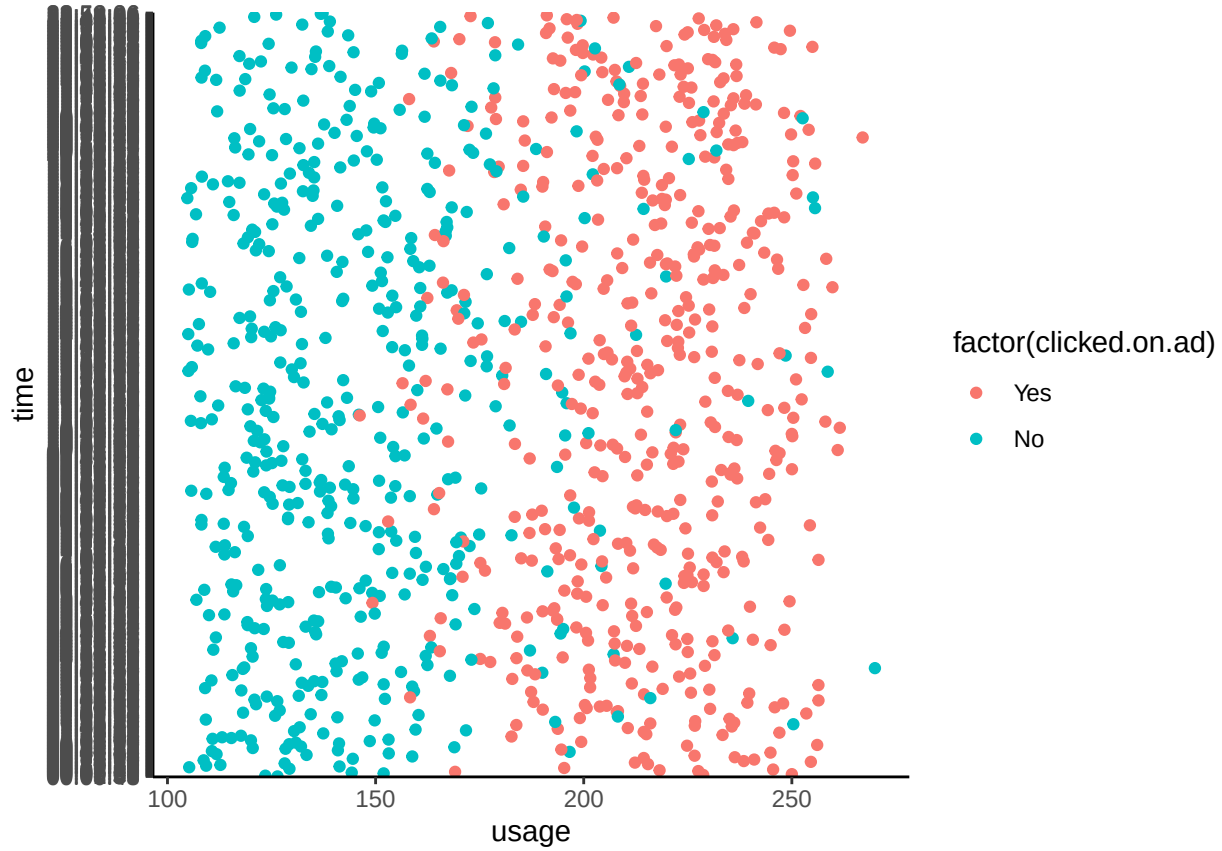

```
##scatter plot for age and usage
#---
#
ggplot(final_df, aes(x=age , y = usage )) + geom_point(aes(colour= factor(`clicked.on.ad`)))
```




```
##scatter plot for age and time
#---
#
ggplot(final_df, aes(x=age , y = time )) + geom_point(aes(colour= factor(`clicked.on.ad`)))
```




```
##scatter plot for usage against time
#---
#
ggplot(final_df, aes(x=usage , y = time )) + geom_point(aes(colour= factor(`clicked.on.ad`)))
```



Conclusion

From the above scatter plot we deduce the following:

1. The scatter plot for income against age shows that majority of the people who failed to click the ads were high income earners with age between 20 and 40 years.

2.The scatter plot for internet usage against income shows that individuals who spend over 200 minutes online are less likely to click an ad.

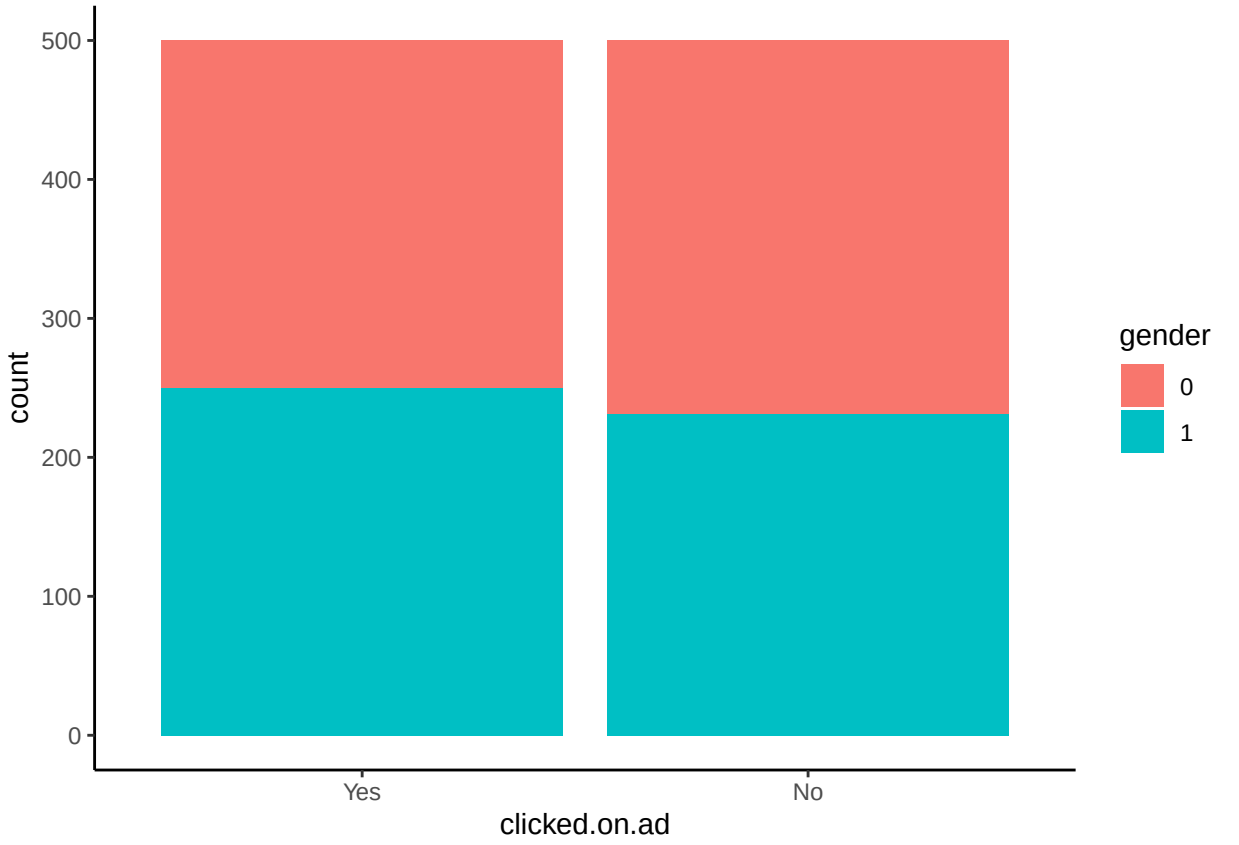
3. The scatter plot for Age against time shows that younger individuals spend alot of tme online and are less tolerant to ads as compared to individuals with age above 35 years.

4. Finally, we observed from the scatter plot for time against usage that the more time one spends online the more the usage.

We continue to Explore which individuals are more likely to click an ad using other visualizations other than scatter plot

```
##Who is likely to click on an ad, female or male?
#---
#
library(ggplot2)

## we use stacked bar chart
#---
```

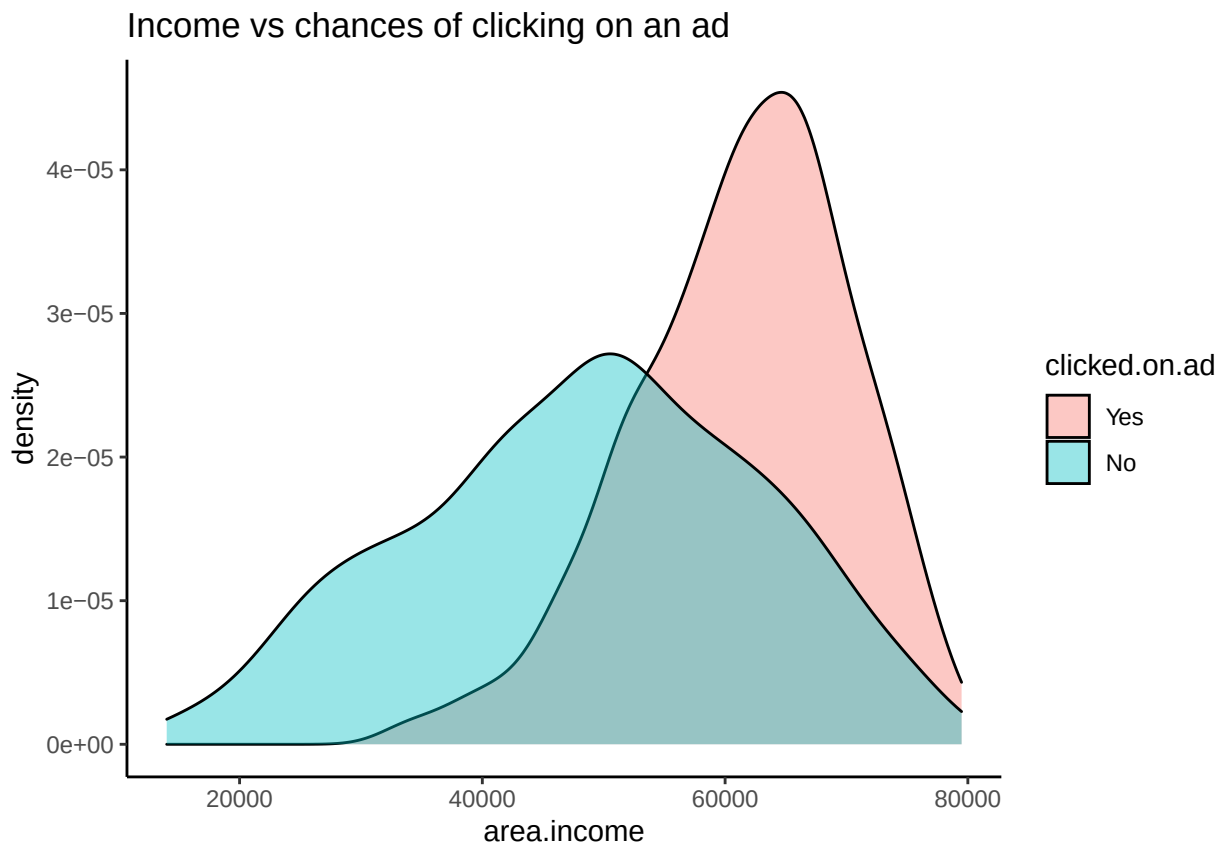
```
#
ggplot(final_df, aes(x = clicked.on.ad,  fill = gender)) + geom_bar(position = "stack")
```



Conclusion

From the above stacked bar chart we observed that Female (where male=0) are more likely to click an ad as compared to male individuals.

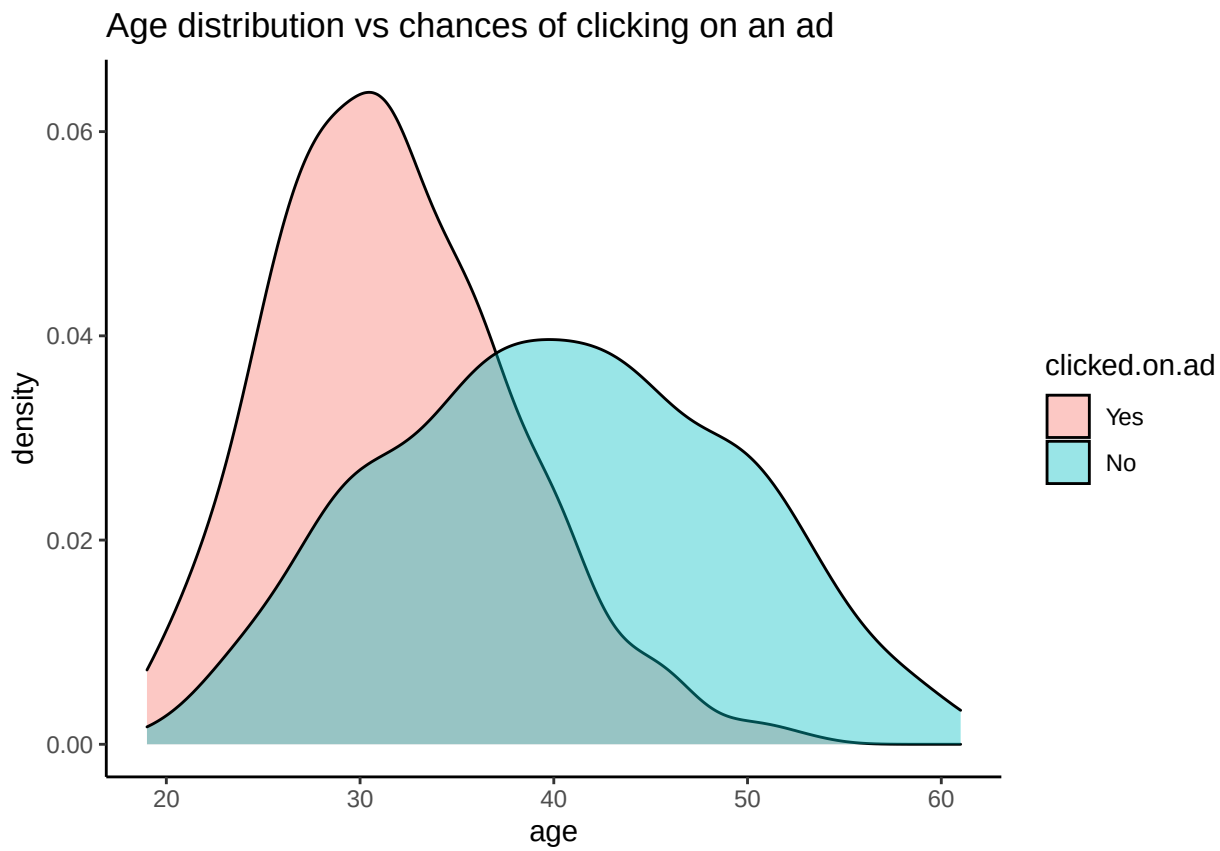
```
##Income class and it's relationship to clicking an ad
#---
#
ggplot(final_df, aes(x = area.income,  fill = clicked.on.ad)) +geom_density(alpha = 0.4) +
  labs(title = "Income vs chances of clicking on an ad")
```



Conclusion

We observed that the income range for people who click on an ad is large as compared to those individuals who failed to click the ad . People from all ranges of income are likely to click on an ad but most of the group with an income of above 50000 are less likely to click on an ad.

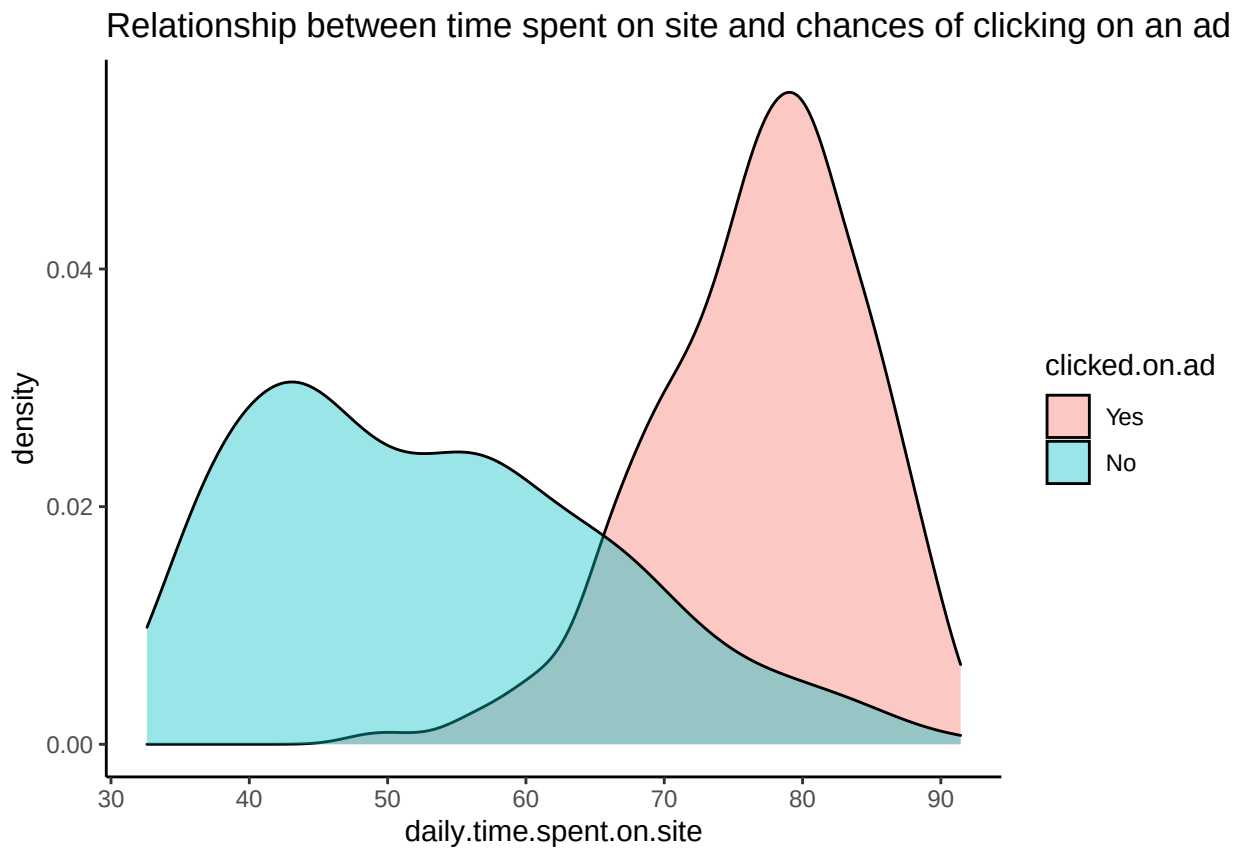
```
##Age and it's relationship to clicking an ad
#---
#
ggplot(final_df, aes(x = age,  fill = clicked.on.ad)) +geom_density(alpha = 0.4) +
  labs(title = "Age distribution vs chances of clicking on an ad")
```



Conclusion

We observed that People above age 35 are more likely to click the ad.

```
## Time spent on site and it's relationship to clicking an ad
ggplot(final_df, aes(x = daily.time.spent.on.site,  fill = clicked.on.ad)) +
  geom_density(alpha = 0.4) +
  labs(title = "Relationship between time spent on site and chances of clicking on an ad")
```



Conclusion

We observed that individuals who spend more time online are less likely to click an add as compared to those who spend less time.

We plot correlation matrix for numerical variables

```
##Selecting the Numerical Variables of the dataset
#---
#
corr <- select(ad , age , 'area.income', `clicked.on.ad`,
               `daily.internet.usage`, `daily.time.spent.on.site`, gender )
```

1.6 Challenging the solution

For better understanding of potential individuals who are more likely to click an ad and also factors that mostly contribute to clicking these ad , modelling predictions should be performed.

1.7 Recommendations

Other than ads the Kenyan entrepreneur should Come up with a better way of advertising the online cryptography course to high income earners and also to those individuals who spend alot of time online.