# Independent Project - Week 12

Kiprop Amos

2022-05-27

## 1. Research question

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.

## 2. Success criteria

The individuals most likely to click on her advertisements are correctly identified

## 3. Research Methodology

- Defining the research questions and work plan
- Loading the dataset
- Previewing the dataset
- Cleaning the dataset which will entail dealing with outliers, duplicates and missing values appropriately
- Performing Univariate, bivariate and multivariate analysis on the dataset
- Discussing the obtained results
- Concluding based on the findings of the research
- Providing recomendations based on the conclusions arrived at

## 4. Understanding the data provided

The dataset that shall be used shall be an advertising dataset that contains a total of 10 features.

A) Age- The age of the individual that clicked the ad

B) Daily Time Spent on Site - The average time an individual spends on the site

C) Area Income - The average income of the area from which the ad was clicked

D) Daily Internet Usage - The daily internet usage information for the area in which the ad was clicked

E) Ad Topic Line - The topic line of the advertisement

F) City - The city from where the ad was clicked

G) Male - The gender of the individual that clicked the add (0- Female, 1- Male)

H) Country - The country from which the add was clicked

I) Timestamp - The time that the ad was clicked

J) Clicked on Add - Contains information whether the individual clicked on the ad or not (0 - Did not click on add, 1 - Clicked on the add)

## 5.

# Loading and previewing the dataset

### Loading the dataset

```
# Reading the advertisement dataset
#
ad_dataset <- read.csv("http://bit.ly/IPAdvertisingData")

# Previewing the first six records of the dataset
#
head(ad_dataset)
```

```
##   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
```

## Previewing the dataset

Here the structure/shape of the dataset, the data types of the various attributes shall be investigated

```
# view the number of rows and columns in the dataset
#
dim(ad_dataset)
```

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

The data set has a total of 1000 records and 10 attributes/columns.

```
# Previewing the structure of the ad dataset
#
str(ad_dataset)
```

```
## '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 are three datatypes in the data set: Number(num), Integer(int), and Character(chr). All attributes have appropriate datatypes excluding the male and clicked on ad columns. These are labelled as integers and are factors. They take only two values (1 or 0)

```
# Loading the library stringr and dplyr
#
library(stringr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Creating a new column that counts the number of words per ad topic lin
#
ad_dataset <- ad_dataset %>%
  mutate(word.Counter = str_count(ad_dataset$Ad.Topic.Line, pattern = "\\w+"))
```

```
# Loading the countrycode library
library(countrycode)

# Grouping the countries according to continent
ad_dataset$continent <- countrycode(sourcevar = ad_dataset[, "Country"],
                            origin = "country.name",
                            destination = "continent")
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : Some va
```

Two columns are being drop (Ad.Topic.Line and Country):

- The ad topic line is dropped since with sentence word counter column It ceases adding value to the study
- The country column is dropped since with the continent data the Country column becomes redundant

```
# Specifying which columns are to be dropped
#
drops <- c("Ad.Topic.Line", "Country")

# Droping the specified column
#
ad_dataset <- ad_dataset[ , !(names(ad_dataset) %in% drops)]

# Printing out the first six records of the dataset
#
head(ad_dataset)
```

```
##   Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage           City
## 1                    68.95  35    61833.90               256.09    Wrightburgh
## 2                    80.23  31    68441.85               193.77      West Jodi
## 3                    69.47  26    59785.94               236.50       Davidton
## 4                    74.15  29    54806.18               245.89 West Terrifurt
## 5                    68.37  35    73889.99               225.58   South Manuel
## 6                    59.99  23    59761.56               226.74      Jamieberg
##   Male            Timestamp Clicked.on.Ad word.Counter continent
## 1    0 2016-03-27 00:53:11             0            3    Africa
## 2    1 2016-04-04 01:39:02             0            3   Oceania
## 3    0 2016-03-13 20:35:42             0            5    Europe
## 4    1 2016-01-10 02:31:19             0            5    Europe
## 5    0 2016-06-03 03:36:18             0            3    Europe
## 6    1 2016-05-19 14:30:17             0            4    Europe
```

The ad topic line and country columns have been successfully dropped.

```
# converting to datetime object
#
ad_dataset[['Timestamp']] <- as.POSIXct(ad_dataset[['Timestamp']],
                                format = "%Y-%m-%d %H:%M:%S")
```

```
# Converting the attribute male from integer to factor
#
as.factor(ad_dataset$Male) -> ad_dataset$Male

# Converting the attribute clicked.on.ad frOm integer to factor
#
as.factor(ad_dataset$Clicked.on.Ad) -> ad_dataset$Clicked.on.Ad

# Converting the attribute word_counter frm integer to factor
#
as.factor(ad_dataset$word.Counter) -> ad_dataset$word.Counter

# Check the structure structure after reassigning the data types
str(ad_dataset)
```

```
## '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 ...
##  $ 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 ...
##  $ Timestamp               : POSIXct, format: "2016-03-27 00:53:11" "2016-04-04 01:39:02" ...
##  $ Clicked.on.Ad           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
##  $ word.Counter            : Factor w/ 5 levels "3","4","5","6",..: 1 1 3 3 1 2 1 1 1 1 ...
##  $ continent               : chr  "Africa" "Oceania" "Europe" "Europe" ...
```

All columns now have appropriate data types. We have numerical, factor, character and POSIXct(datetime datatypes)

```
# Establish the data set class
#
class(ad_dataset)
```

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

The advertisement dataset is a data frame

# Cleaning Dataset

## Dealing with missing values

```
# Checking the number of missing values per column in the data set
#
colSums(is.na(ad_dataset))
```

```
## Daily.Time.Spent.on.Site                      Age              Area.Income
##                        0                        0                        0
##     Daily.Internet.Usage                     City                     Male
```

```
##                         0                         0                         0
##                 Timestamp            Clicked.on.Ad             word.Counter
##                         0                         0                         0
##                 continent
##                        35
```

The dataset has no missing values in any of the attributes

## Checking for duplicate records

```
# finding the duplicated rows in the data set df and assign to a variable duplicated_rows below
# ---
#
duplicated_rows = ad_dataset[duplicated(ad_dataset),]

# Printing out the duplicated rows
duplicated_rows
```

```
##  [1] Daily.Time.Spent.on.Site Age                       Area.Income
##  [4] Daily.Internet.Usage     City                      Male
##  [7] Timestamp                Clicked.on.Ad             word.Counter
## [10] continent
## <0 rows> (or 0-length row.names)
```

The advertisement data set has no duplicate records

## Checking for outliers in the numeric data

```
#  Identifying the numeric class in the data and evaluating if there are any
#  outliers
#
num_cols <- unlist(lapply(ad_dataset, is.numeric)) # Identify numeric columns

# Printing out num_cols
#
num_cols
```

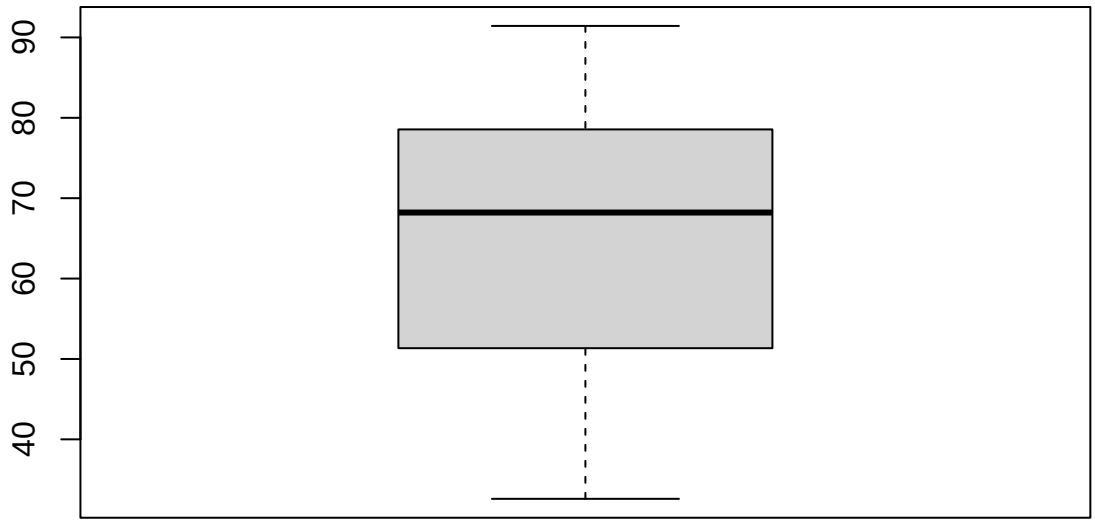
```
## Daily.Time.Spent.on.Site                       Age             Area.Income
##                     TRUE                      TRUE                     TRUE
##     Daily.Internet.Usage                      City                     Male
##                     TRUE                     FALSE                    FALSE
##                Timestamp            Clicked.on.Ad             word.Counter
##                    FALSE                     FALSE                    FALSE
##                continent
##                    FALSE
```

```
# Subset numeric columns of data
#
data_num <- ad_dataset[ , num_cols]

# Printing the subset to RStudio console
#
head(data_num)
```

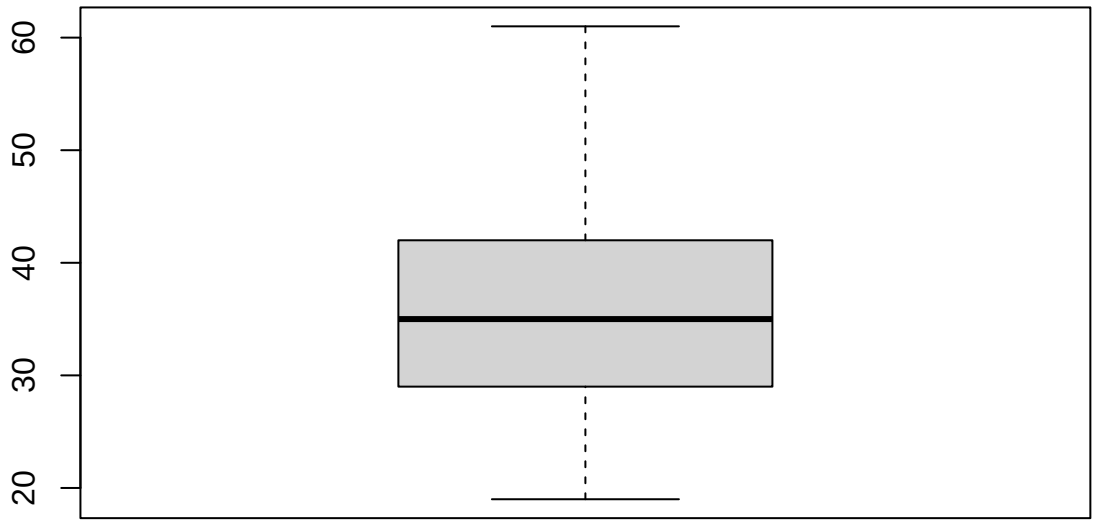
```
##   Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage
## 1                    68.95  35    61833.90               256.09
## 2                    80.23  31    68441.85               193.77
## 3                    69.47  26    59785.94               236.50
## 4                    74.15  29    54806.18               245.89
## 5                    68.37  35    73889.99               225.58
## 6                    59.99  23    59761.56               226.74
```

There are four columns with numeric data; Daily.Time.Spent.on.Site, Age, Area.Income, and Daily.Internet.Usage

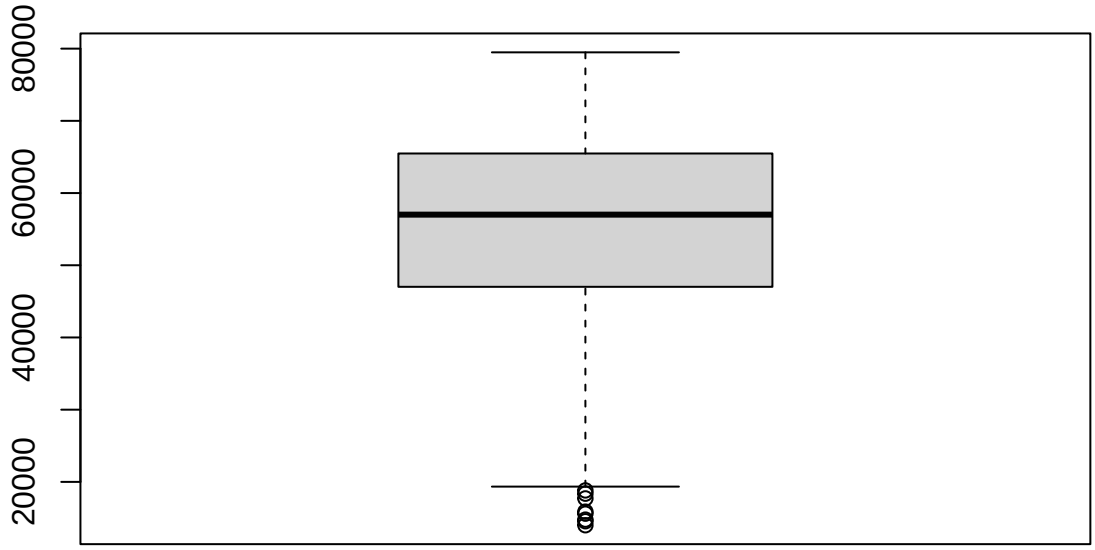
```
# Box plot of the four numeric columns to check for outliers

# Box plot of Daily.Time.Spent.on.Site column
#
boxplot(data_num$Daily.Time.Spent.on.Site)
```



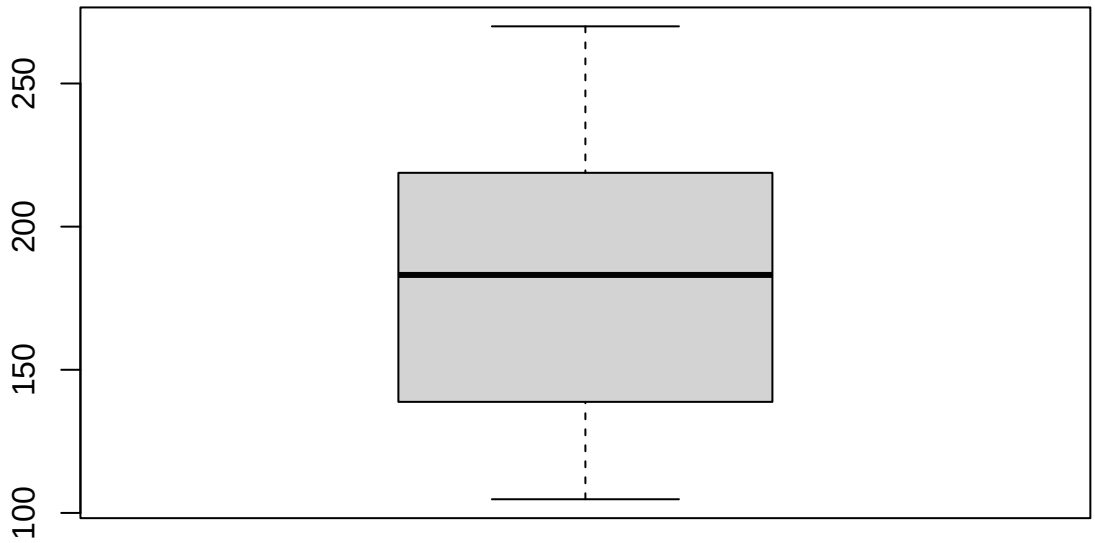
```r
# Box plot of age column
#
boxplot(data_num$Age)
```




```r
# Box plot of of area income column
#
boxplot(data_num$Area.Income)
```




```
# Box plot of daily internet usage
#
boxplot(data_num$Daily.Internet.Usage)
```




Outliers were observed only in the attribute containing area income information. This is expected due to the great disparity in development and GDP levels for the different countries globally.

# Univariate analysis

## Measures of central tendency

### Mean

```
# Get Mean of the multiple numeric columns in ad dataset
#
colMeans(ad_dataset[sapply(ad_dataset, is.numeric)])
```

```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##                  65.0002                  36.0090               55000.0001
##     Daily.Internet.Usage
##                180.0001
```

Individuals visiting the site had an average age of 36.0090 years On average an individual spent 65.0002 seconds on the site On avearage an individual used 180.0001 mbs of data The average area income of the surveyed individuals was 55000.0001 US$

```r
# Getting the means of the various numeric attributes based on whether the
# add was clicked (ad.clicked.on = 1)

# Selecting records that had the ad clicked (ad.clicked.on = 1)
#
ad_clicked <- ad_dataset[ad_dataset$Clicked.on.Ad == '1', ]

# Get Mean of the multiple numeric columns when the add was clicked on
#
colMeans(ad_clicked[sapply(ad_clicked, is.numeric)])
```

```
## Daily.Time.Spent.on.Site                  Age              Area.Income
##                 53.14578             40.33400             48614.41374
##     Daily.Internet.Usage
##                145.48646
```

-> The average age of internet user that clicked on add was 40.334 years -> The average time an individual who clicked on a add spent on the site was 53.14578 seconds -> The average area income of the individual that clicked the add was 48614.41374 US$ -> The average daily internet usage of the individual that clicked the add was 145.48646 mbs

```r
# Getting the means of the various numeric attributes based on whether the
# add was not clicked (ad.clicked.on = 0)

# Selecting records that had the ad not clicked (ad.clicked.on = 0)
#
ad_not_clicked <- ad_dataset[ad_dataset$Clicked.on.Ad == '0', ]

# Get Mean of the multiple numeric columns when the add was not clicked on
#
colMeans(ad_not_clicked[sapply(ad_not_clicked, is.numeric)])
```

```
## Daily.Time.Spent.on.Site                  Age              Area.Income
##                 76.85462             31.68400             61385.58642
##     Daily.Internet.Usage
##                214.51374
```

-> The average age of internet user that did not click on an ad was 31.68400 years -> The average time an individual who did not click on an ad spent on the site was 76.85462 seconds -> The average area income of the individual that did not click on an ad was 61385.58642 US$ -> The average daily internet usage of the individual that did not click on an ad was 214.51374 mbs

### Median

```r
# Finding the median of all columns in numeric dataframe num_data
#
apply(data_num, 2,median)
```

```
## Daily.Time.Spent.on.Site                  Age              Area.Income
##                   68.215               35.000                57012.300
##     Daily.Internet.Usage
##                  183.130
```

-> The mid time spent on site by individuals on the site spent was 68.215 seconds -> The mid age of individuals on the site was 35 years -> The mid area income of individuals on the site was 57012.300 US$ -> The mid daily internet usage by individuals on the site was 183.130 mbs

```
# Getting the medians of the various numeric attributes based on whether
# the add was clicked (ad.clicked.on = 1)

#  Isolating the numeric class in the ad_clicked dataset
#
num_cols <- unlist(lapply(ad_clicked, is.numeric)) # Identify numeric columns

# Subset numeric columns of data
#
data_num1 <- ad_clicked[ , num_cols]

# Get Medians of the multiple numeric columns when the add was clicked on
#
apply(data_num1, 2,median)
```

```
## Daily.Time.Spent.on.Site                     Age             Area.Income
##                    51.53                   40.00                49417.26
##     Daily.Internet.Usage
##                   138.79
```

-> The mid age of individuals that clicked the ad was 40.0 years -> The mid time spent on site by individuals that clicked the ad was 51.53 seconds -> The mid area income of the individuals that clicked the ad was 49417.26 US$ -> The mid daily internet usage of individuals that clicked the ad was 138.79 mbs

```
# Getting the medians of the various numeric attributes based on whether the
# add was not clicked (ad.clicked.on = 0)

#  Isolating the numeric class in the ad_clicked dataset
# Identify numeric columns
#
num_cols <- unlist(lapply(ad_not_clicked, is.numeric))

# Subset numeric columns of data
#
data_num2 <- ad_not_clicked[ , num_cols]

# Get Medians of the multiple numeric columns when the add was not clicked on
#
apply(data_num2, 2,median)
```

```
## Daily.Time.Spent.on.Site                     Age             Area.Income
##                   77.650                  31.000               62275.405
##     Daily.Internet.Usage
##                  216.365
```

-> The mid age of individuals that did not click the ad was 31.0 years -> The mid time spent on site by individuals that did not click the ad was 77.65 seconds -> The mid area income of the individuals that did not click the ad was 62775.405 US$ -> The mid daily internet usage of individuals that did not click the ad was 216.365 mbs

**Mode**

```
# Creating the mode function that will perform our mode operation for us
# ---
getmode <- function(v) {
uniqv <- unique(v)
uniqv[which.max(tabulate(match(v, uniqv)))]
}

# iterating over all the columns of the
# dataframe
#
for (i in 1:ncol(data_num)){

  # calculating mode of ith column
  mod_val <- getmode(data_num[,i])
  cat(i, ": ",mod_val,"\n")
}
```

```
## 1 :  62.26
## 2 :  31
## 3 :  61833.9
## 4 :  167.22
```

-> Most of the individuals that were on the site were 31.0 years

-> Most individuals that were on the site spent 62.26 seconds on the site

-> Most of the individuals that were on the site came from areas with income of 61833.9 US$

-> Most of the individuals that were on the site had a daily internet usage of 167.22 mbs

```
# Get the mode of values of the records that showed the ad was clicked on
#

# iterating over all the columns of the
# dataframe
for (i in 1:ncol(data_num1)){

  # calculating mode of ith column
  mod_val <- getmode(data_num1[,i])
  cat(i, ": ",mod_val,"\n")
}
```

```
## 1 :  75.55
## 2 :  45
## 3 :  24593.33
## 4 :  167.22
```

-> Most of the individuals that clicked the ad were 45 years

-> Most individuals that clicked the ad spent 75.55 seconds on the site

-> Most of the individuals that clicked the ad came from areas with income of 24593.33 US$

-> Most of the individuals that clicked the ad had a daily internet usage of 167.22 mbs

```r
# Get the mode of values of the records that showed the ad were not
# clicked on

# iterating over all the columns of the
# dataframe
for (i in 1:ncol(data_num2)){

  # calculating mode of ith column
  mod_val <- getmode(data_num2[,i])
  cat(i, ": ",mod_val,"\n")
}
```

```
## 1 :  77.05
## 2 :  31
## 3 :  61833.9
## 4 :  235.28
```

-> Most of the individuals that did not click the ad were 31.0 years

-> Most individuals that did not click the ad spent 77.05 seconds on the site

-> Most of the individuals that did not click the ad came from areas with income of 61833.9 US$

-> Most of the individuals that did not click the ad had a daily internet usage of 235.28 mbs

## Measures of dispersion

### Range

```r
# Range of age, internet usage, time spent on site and average area income
#
for (i in 1:ncol(data_num)){

  # calculating mode of ith column
  range_val <- range(data_num[,i])
  cat(i, ": ",range_val,"\n")
}
```

```
## 1 :  32.6 91.43
## 2 :  19 61
## 3 :  13996.5 79484.8
## 4 :  104.78 269.96
```

-> The minimum spent by a user on the site was 32.6 seconds the maximum time was 91.43 seconds

-> The minimum age of a user of the was site was 19 years the maximum age 61 years

-> The minimum area income of the site user was 13996.5 US$ the maximum area income 79484.8

-> The minimum daily internet usage by a site visitor was 104.78 mbs while the maximum was 269.96 mbs

```r
# Range of age, internet usage, time spent on site and average area income
# for records that indicate ad was clicked
#
for (i in 1:ncol(data_num1)){

  # calculating mode of ith column
  range_val <- range(data_num1[,i])
  cat(i, ": ",range_val,"\n")
}
```

```
## 1 :  32.6 91.37
## 2 :  19 61
## 3 :  13996.5 78520.99
## 4 :  104.78 269.96
```

-> The minimum time spent by a user who clicked the ad on the site was 32.6 seconds the maximum time was 91.43 seconds

-> The minimum age of a user of the site who clicked the ad was 19 years the maximum age 61 years

-> The minimum area income of the site user who clicked the ad was 13996.5 US$ the maximum area income 78520.99

-> The minimum daily internet usage by a site visitor who clicked the ad was 104.78 mbs while the maximum was 269.96 mbs

```r
# Range of age, internet usage, time spent on site and average area income
# for records that indicate ad was not clicked
#
for (i in 1:ncol(data_num2)){

  # calculating mode of ith column
  range_val <- range(data_num2[,i])
  cat(i, ": ",range_val,"\n")
}
```

```
## 1 :  48.22 91.43
## 2 :  19 53
## 3 :  33239.2 79484.8
## 4 :  146.19 267.01
```

-> The minimum time spent by a user who did not click the ad on the site was 48.22 seconds the maximum time was 91.43 seconds

-> The minimum age of a user of the site did not click the ad was 19 years the maximum age 53 years

-> The minimum area income of the site user did not click the ad was 33239.2 US$ the maximum area income 79484.8

-> The minimum daily internet usage by a site visitor did not click the ad was 146.19 mbs while the maximum was 267.01 mbs

**Quantile range**

```r
# quantile Range of age, internet usage, time spent on site and average area income
# for records that indicate ad was not clicked
#
for (i in 1:ncol(data_num)){

  # calculating mode of ith column
  quantile_val <- quantile(data_num[,i])
  cat(i, ": ",quantile_val,"\n")
}
```

```
## 1 :  32.6 51.36 68.215 78.5475 91.43
## 2 :  19 29 35 42 61
## 3 :  13996.5 47031.8 57012.3 65470.64 79484.8
## 4 :  104.78 138.83 183.13 218.7925 269.96
```

The 0% 25% 50% 75% 100% respectively are;

time spent on site -> (32.6 51.36 68.215 78.5475 91.43) seconds

age -> (19 29 35 42 61) years

average area income -> (13996.5 47031.8 57012.3 65470.64 79484.8) US$

internet usage -> (104.78 138.83 183.13 218.7925 269.96)mbs

**Standard Deviation**

```r
# Standard Deviation of age, internet usage, time spent on site
# and average area income
#
for (i in 1:ncol(data_num)){

  # calculating Standard Deviation of ith column
  sd_val <- sd(data_num[,i])
  cat(i, ": ",sd_val,"\n")
}
```

```
## 1 :  15.85361
## 2 :  8.785562
## 3 :  13414.63
## 4 :  43.90234
```

For all site users:

-> Time spent on site by a user had a Standard Deviation of 15.85361 secs

-> Age of the site users had a Standard Deviation of 8.785562 years

-> The area income had a Standard Deviation of 13414.63 US$

-> The daily internet usage had a Standard Deviation of 43.90234 mbs

```r
# Standard Deviation of age, internet usage, time spent on site
# and average area income

# for records that indicate ad was clicked
#
for (i in 1:ncol(data_num1)){

  # calculating Standard Deviation of ith column
  sd_val <- sd(data_num1[,i])
  cat(i, ": ",sd_val,"\n")
}
```

```
## 1 :  12.82209
## 2 :  8.856598
## 3 :  14116.24
## 4 :  30.02583
```

For site users that clicked the ad:

-> Time spent on site by a user had a Standard Deviation of 12.82209 secs

-> Age of the site users had a Standard Deviation of 8.856598 yrs

-> The area income had a Standard Deviation of 14116.24 US$

-> The daily internet usage had a Standard Deviation of 30.02583 mbs

```r
# Standard Deviation of age, internet usage, time spent on site
# and average area income

# for records that indicate ad was not clicked
#
for (i in 1:ncol(data_num2)){

  # calculating Standard Deviation of ith column
  sd_val <- sd(data_num2[,i])
  cat(i, ": ",sd_val,"\n")
}
```

```
## 1 :  7.560031
## 2 :  6.212998
## 3 :  8904.06
## 4 :  23.87438
```

For site users that clicked the ad:

-> Time spent on site by a user had a Standard Deviation of 7.560031

-> Age of the site users had a Standard Deviation of 6.212998 yrs

-> The area income had a Standard Deviation of 8904.06 US$

-> The daily internet usage had a Standard Deviation of 23.87438 mbs

**Variance**

```r
# Variance of age, internet usage, time spent on site
# and average area income
#
for (i in 1:ncol(data_num)){

  # calculating variance of ith column
  var_val <- var(data_num[,i])
  cat(i, ": ",var_val,"\n")
}
```

```
## 1 :  251.3371
## 2 :  77.18611
## 3 :  179952406
## 4 :  1927.415
```

For all site users:

-> Time spent on site by a user had a variation of 251.3371

-> Age of the site users had a variation of 77.18611

-> The area income had a variation of 179952406

-> The daily internet usage had a variation of 1927.415

```r
# Variance of age, internet usage, time spent on site
# and average area income

# for records that indicate ad was clicked
#
for (i in 1:ncol(data_num1)){

  # calculating variance of ith column
  var_val <- var(data_num1[,i])
  cat(i, ": ",var_val,"\n")
}
```

```
## 1 :  164.406
## 2 :  78.43932
## 3 :  199268295
## 4 :  901.5502
```

For site users that clicked the ad:

-> Time spent on site by a user had a variation of 164.406

-> Age of the site users had a variation of 78.43932

-> The area income had a variation of 199268295

-> The daily internet usage had a variation of 901.5502

```r
# Variance of age, internet usage, time spent on site
# and average area income

# for records that indicate ad was not clicked
```

```
#
for (i in 1:ncol(data_num2)){

  # calculating variance of ith column
  var_val <- var(data_num2[,i])
  cat(i, ": ",var_val,"\n")
}
```

```
## 1 :  57.15408
## 2 :  38.60135
## 3 :  79282288
## 4 :  569.9861
```

For site users that clicked the ad:

-> Time spent on site by a user had a variation of 57.15408

-> Age of the site users had a variation of 38.60135

-> The area income had a variation of 79282288

-> The daily internet usage had a variation of 569.9861

### Univariate Graphical

Here the non_numeric data shall be visualized to draw insights into the traits of the users that clicked and those that did not click the ads

```
# loading the purrr package from tidyverse
#
library(purrr)

# Selecting only non_numeric data
#
ad_non_numeric <- ad_dataset %>% discard(is.numeric)

# Preview the first six records of non_numeric dataframe
#
head(ad_non_numeric)
```

```
##             City Male           Timestamp Clicked.on.Ad word.Counter continent
## 1    Wrightburgh    0 2016-03-27 00:53:11             0            3    Africa
## 2      West Jodi    1 2016-04-04 01:39:02             0            3   Oceania
## 3       Davidton    0 2016-03-13 20:35:42             0            5    Europe
## 4 West Terrifurt    1 2016-01-10 02:31:19             0            5    Europe
## 5   South Manuel    0 2016-06-03 03:36:18             0            3    Europe
## 6      Jamieberg    1 2016-05-19 14:30:17             0            4    Europe
```

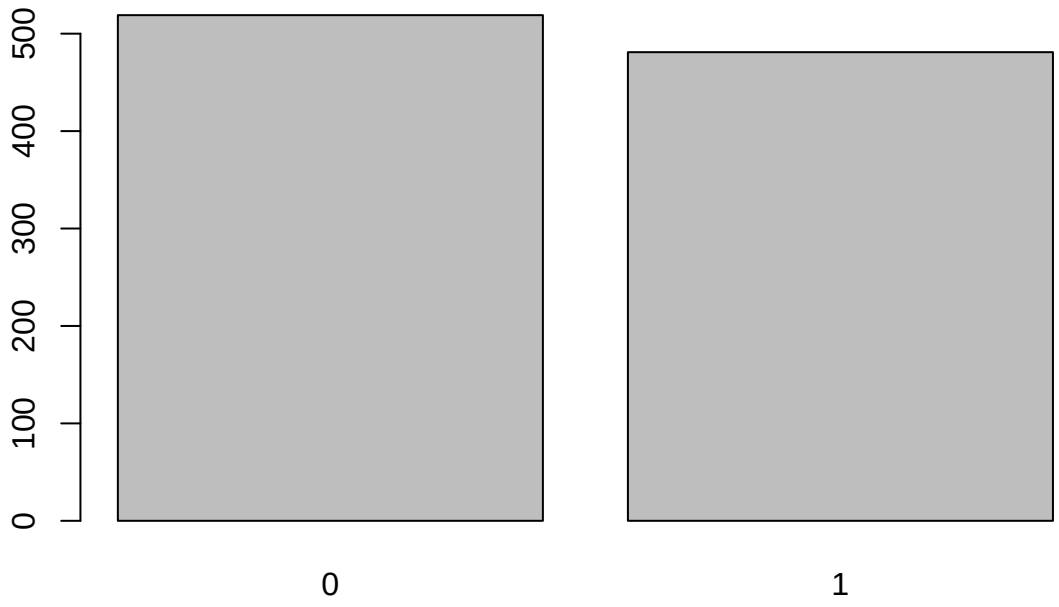
### Value counts of the non numeric data

```
# frequencies of the number of males and females that visited the site
#
gender_freq <- table(ad_dataset$Male)

# Printing out the frequencies
#
gender_freq
```

**Gender**

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

```
# Applying the barplot function to produce its bar graph
# ---
barplot(gender_freq)
```



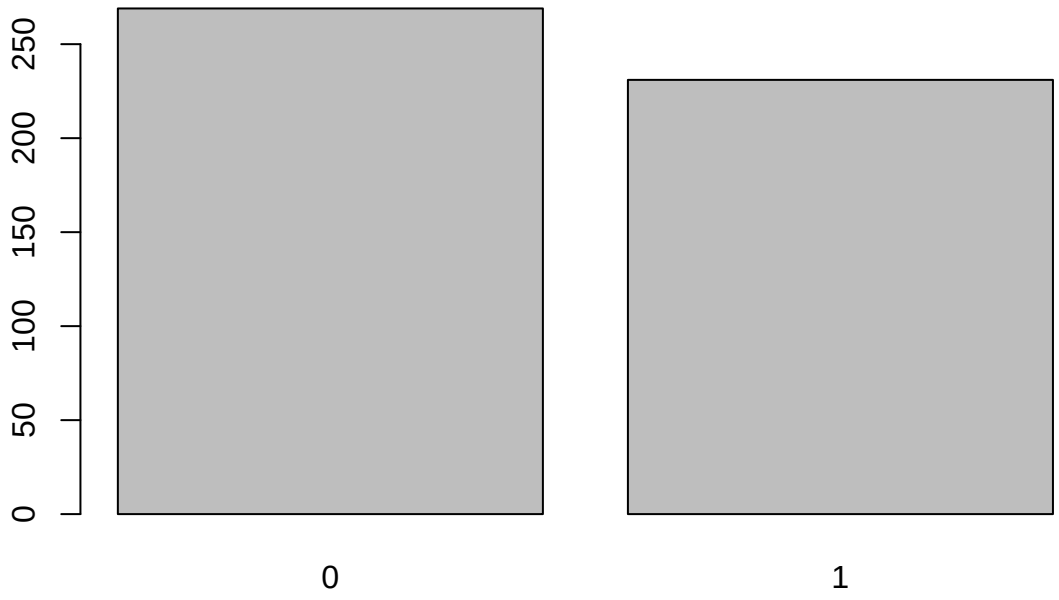

519 of the people that accessed the site were female, 481 were male

```
# frequencies of the number of males and females that clicked
# on the ad
#
gender_freq1 <- table(ad_clicked$Male)
```

```
# Printing out the frequencies
#
gender_freq1
```

```
##
##   0   1
## 269 231
```

```
# Applying the barplot function to produce its bar graph
# ---
barplot(gender_freq1)
```

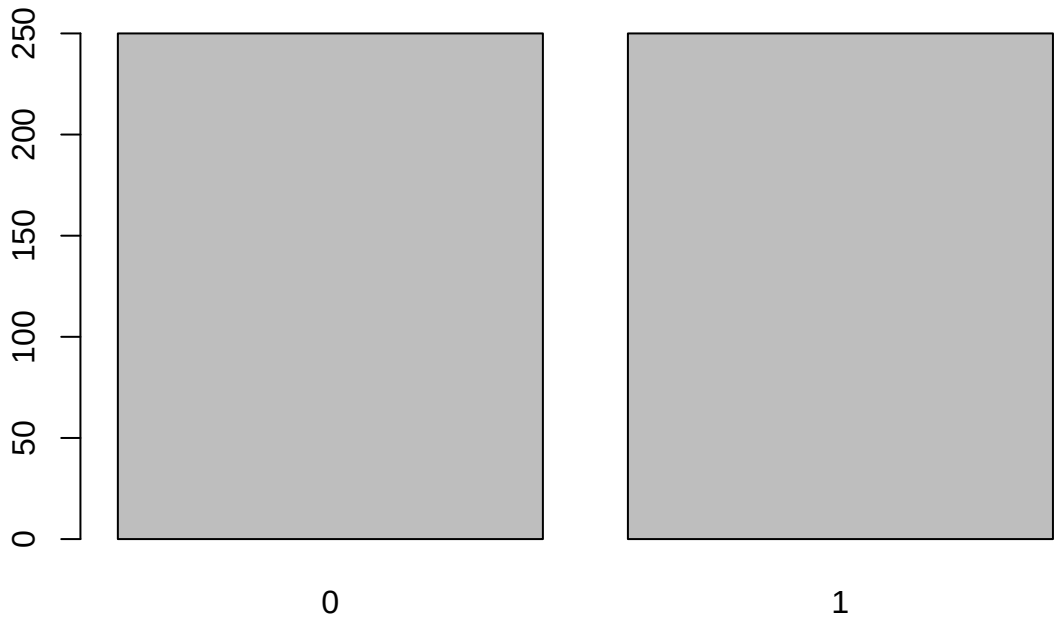



269 of the respondents that clicked the adds were female 231 were male

```
# frequencies of the number of males and females that did not click
# on the ad
#
gender_freq2 <- table(ad_not_clicked$Male)

# Printing out the frequencies
#
gender_freq2
```

```
##
##   0   1
## 250 250
```

```
# Applying the bar plot function to produce its bar graph
# ---
barplot(gender_freq2)
```



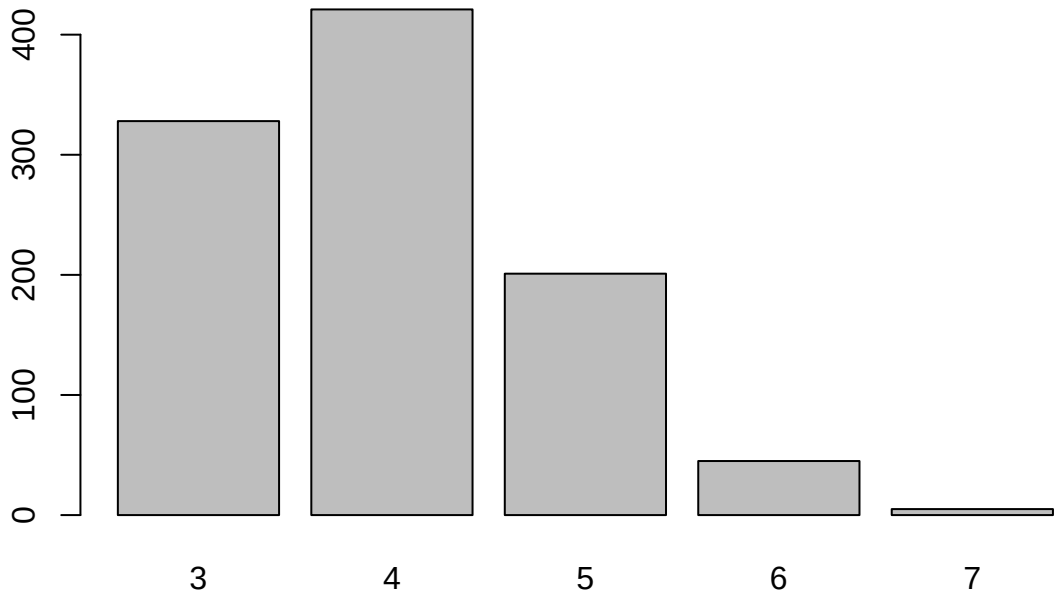

250 of the respondents that clicked the adds were female 250 were male

```
# Frequencies of the word.counter(number of words per ad line) in ad
# Dataset
#
counter_freq <- table(ad_dataset$word.Counter)

# Printing out the frequencies of word counts of different ad topic lines
#
counter_freq
```

**Word counter(Words per ad line)**

```
##
##   3   4   5   6   7
## 328 421 201  45   5
```

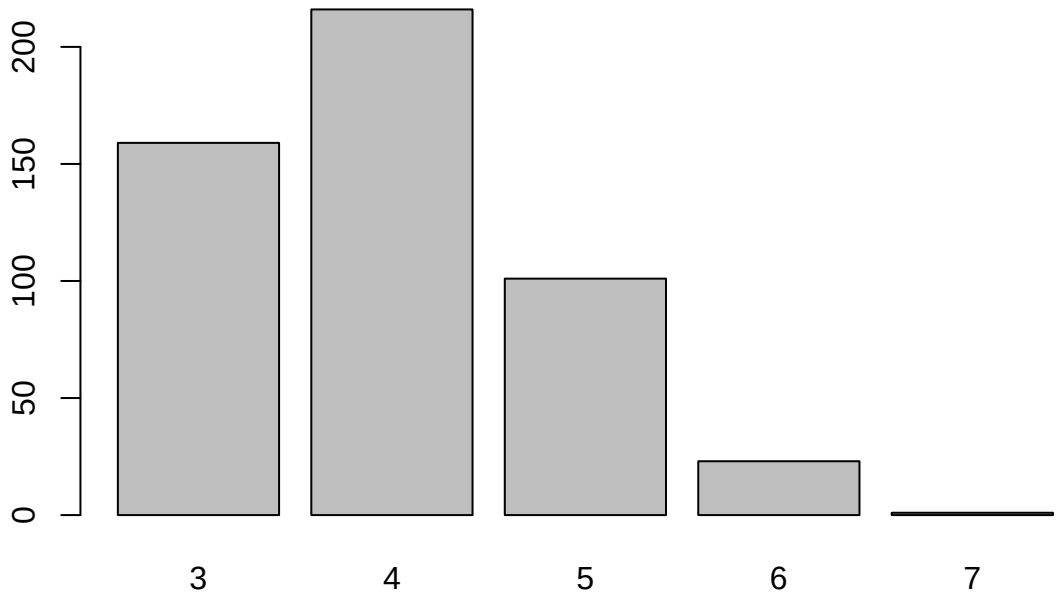
```
# Applying the barplot function to produce its bar graph
# ---
barplot(counter_freq)
```




```
# Frequencies of the word.counter(number of words per ad line) that were
# clicked
#
counter_freq1 <- table(ad_clicked$word.Counter)

# Printing out the frequencies of word counts of different ad topic lines
#
counter_freq1
```

```
##
##   3   4   5   6   7
## 159 216 101  23   1
```

```
# Applying the barplot function to produce its bar graph
# ---
barplot(counter_freq1)
```

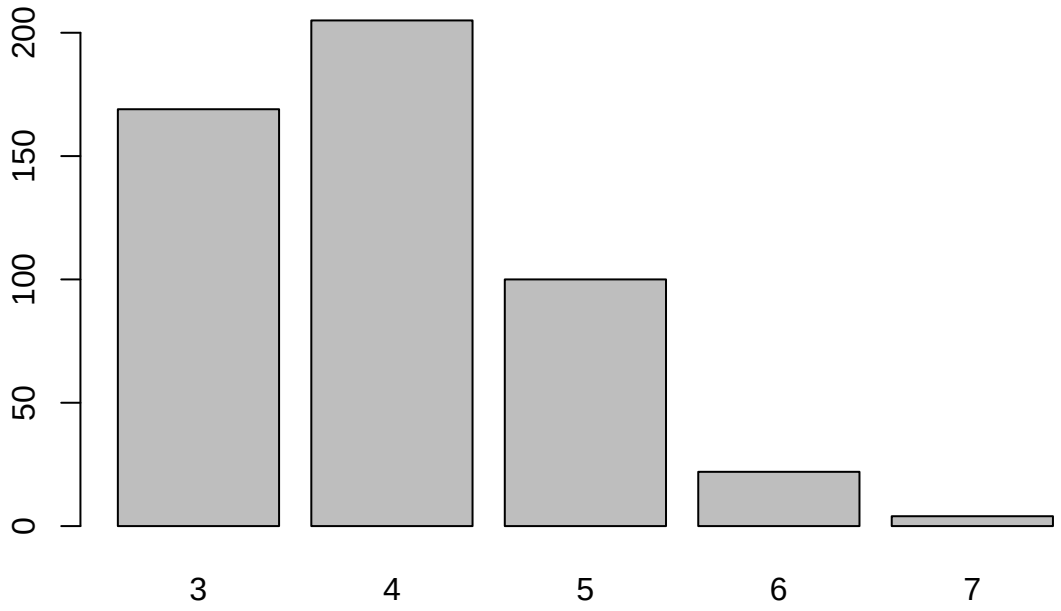



Ad topic lines with 4 words were clicked 206 times Those with 7 lines were clicked only once

```
# Frequencies of the word.counter(number of words per ad line) that were not
# clicked
#
counter_freq2 <- table(ad_not_clicked$word.Counter)

# Printing out the frequencies of word counts of different ad topic lines
#
counter_freq2
```

```
##
##   3   4   5   6   7
## 169 205 100  22   4
```

```
# Applying the barplot function to produce its bar graph
# ---
barplot(counter_freq2)
```



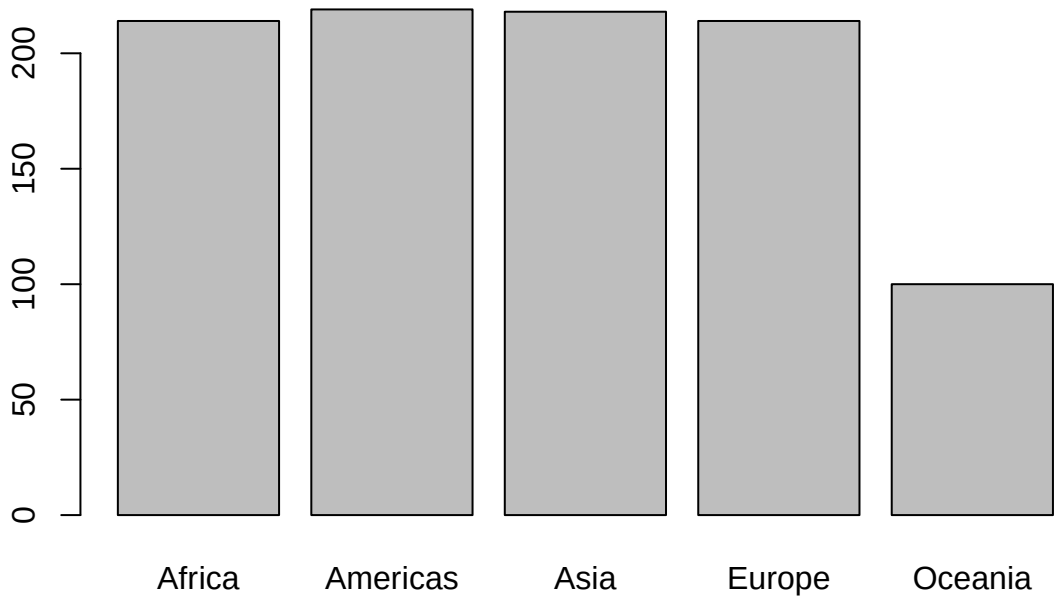

Ad topic lines with 4 words were not clicked 205 times Those with 7 lines were not clicked 4 times

**Continents** Here the ad clicks and ads not clicked from different continents are going to be determined

```
# Frequencies of site visits originating from different regions globally
continent_freq <- table(ad_dataset$continent)

# Printing out the frequencies of the continents by ads ot clicked
#
continent_freq
```

```
##
##   Africa Americas     Asia   Europe  Oceania
##      214      219      218      214      100
```

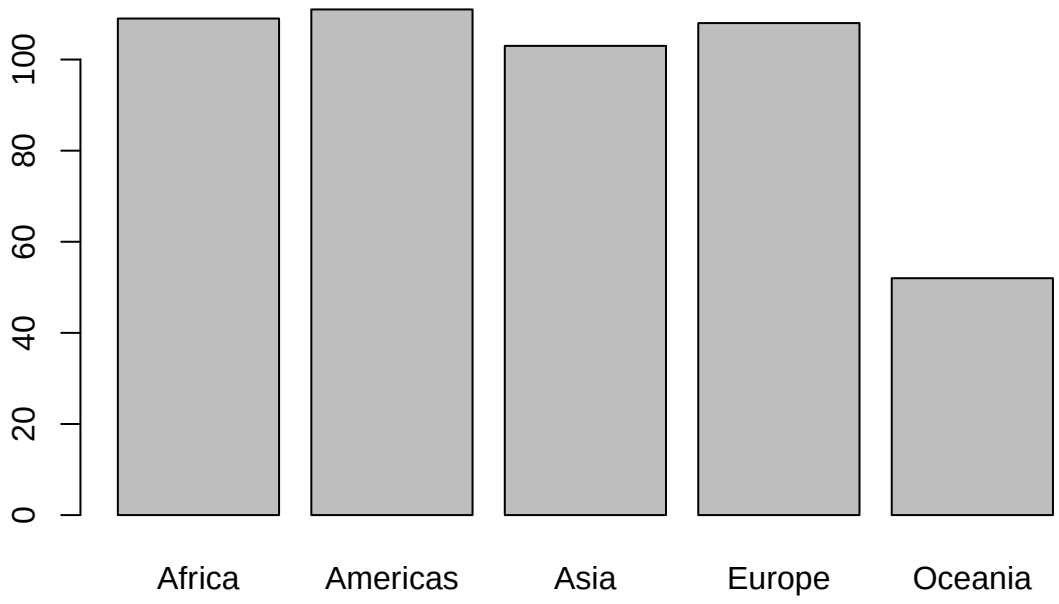
```
# Applying the barplot function to produce its bar graph
# ---
barplot(continent_freq)
```




```
# Frequencies of ad clicks that originated from the different
# continents globally
#
continent_freq1 <- table(ad_clicked$continent)

# Printing out the frequencies of the continents by ads ot clicked
#
continent_freq1
```

```
##
##   Africa Americas     Asia   Europe  Oceania
##      109      111      103      108       52
```

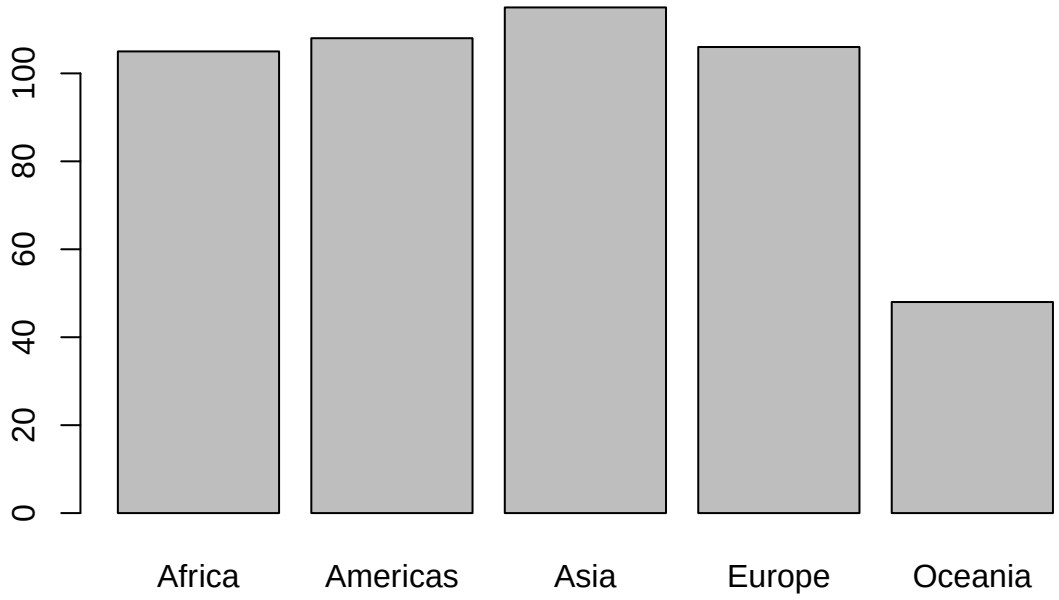
```
# Applying the barplot function to produce its bar graph
# ---
barplot(continent_freq1)
```




```
# Frequencies of ads that were not clicked that originated from the different
# continents globally
#
continent_freq2 <- table(ad_not_clicked$continent)

# Printing out the frequencies of the continents by ads ot clicked
continent_freq2
```

```
##
##   Africa Americas     Asia   Europe  Oceania
##      105      108      115      106       48
```

```
# Applying the bar plot function to produce its bar graph
# ---
barplot(continent_freq2)
```




## Bivariate analysis

### Covariance

Covariance is a statistical representation of the degree to which two variables vary together.Here the relationship between the different numerical data in data Frame shall be calculated

```
# Create Covariance matrix of the numerical data in dataset
#
cov(data_num)
```

```
##                           Daily.Time.Spent.on.Site          Age  Area.Income
## Daily.Time.Spent.on.Site                251.33709    -46.17415     66130.81
## Age                                     -46.17415     77.18611    -21520.93
## Area.Income                           66130.81091 -21520.92580 179952405.95
## Daily.Internet.Usage                    360.99188   -141.63482    198762.53
##                           Daily.Internet.Usage
## Daily.Time.Spent.on.Site             360.9919
## Age                                 -141.6348
## Area.Income                       198762.5315
## Daily.Internet.Usage                1927.4154
```

From the covariance matrix, age varied negatively with all other numerical variables; Daily time spent on site, area income, and daily internet usage.

The other variables have a positive covariance among each other.

### Correlation

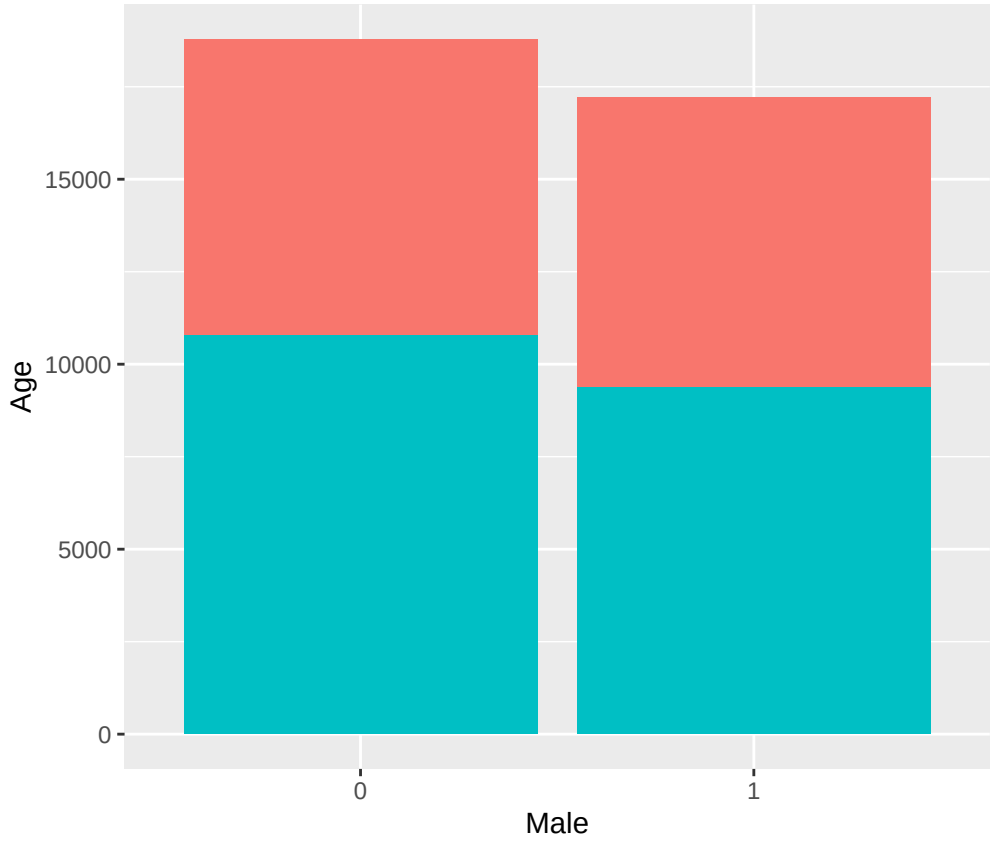
```
# Correlation matrix of numerical data in the ad dataset
#
cor(data_num)
```

```
##                           Daily.Time.Spent.on.Site        Age Area.Income
## Daily.Time.Spent.on.Site                1.0000000 -0.3315133   0.3109544
## Age                                    -0.3315133  1.0000000  -0.1826050
## Area.Income                             0.3109544 -0.1826050   1.0000000
## Daily.Internet.Usage                    0.5186585 -0.3672086   0.3374955
##                           Daily.Internet.Usage
## Daily.Time.Spent.on.Site            0.5186585
## Age                                -0.3672086
## Area.Income                         0.3374955
## Daily.Internet.Usage                1.0000000
```

Age has a negative correlation with the other numerical variables. All other variables positive correlation among each other

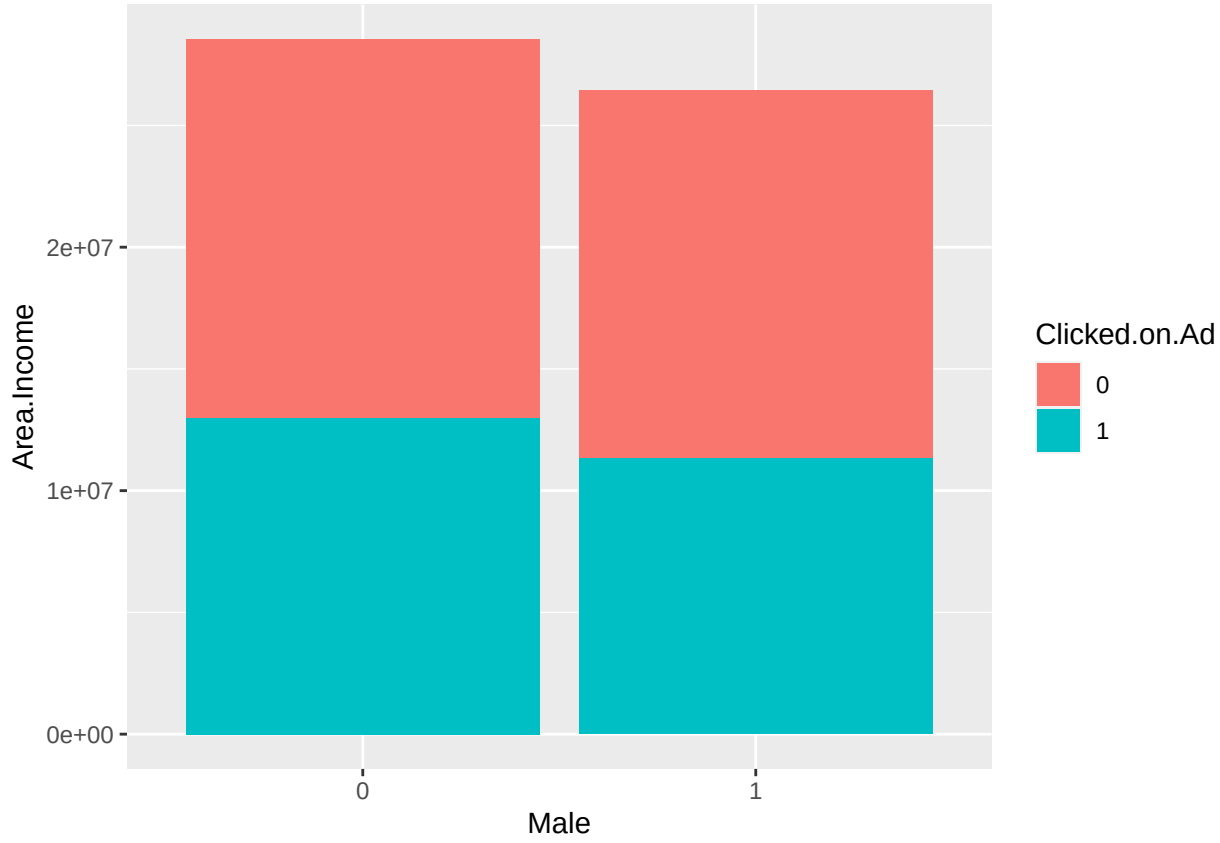
### Bivariate graphical plots

```
# Loading ggplot library
#
library(ggplot2)

# Plotting Stacked column of male(gender) vs age
#
ggplot(ad_dataset, aes(fill=Clicked.on.Ad, y=Age, x=Male)) +
  geom_bar(position='stack', stat='identity')
```

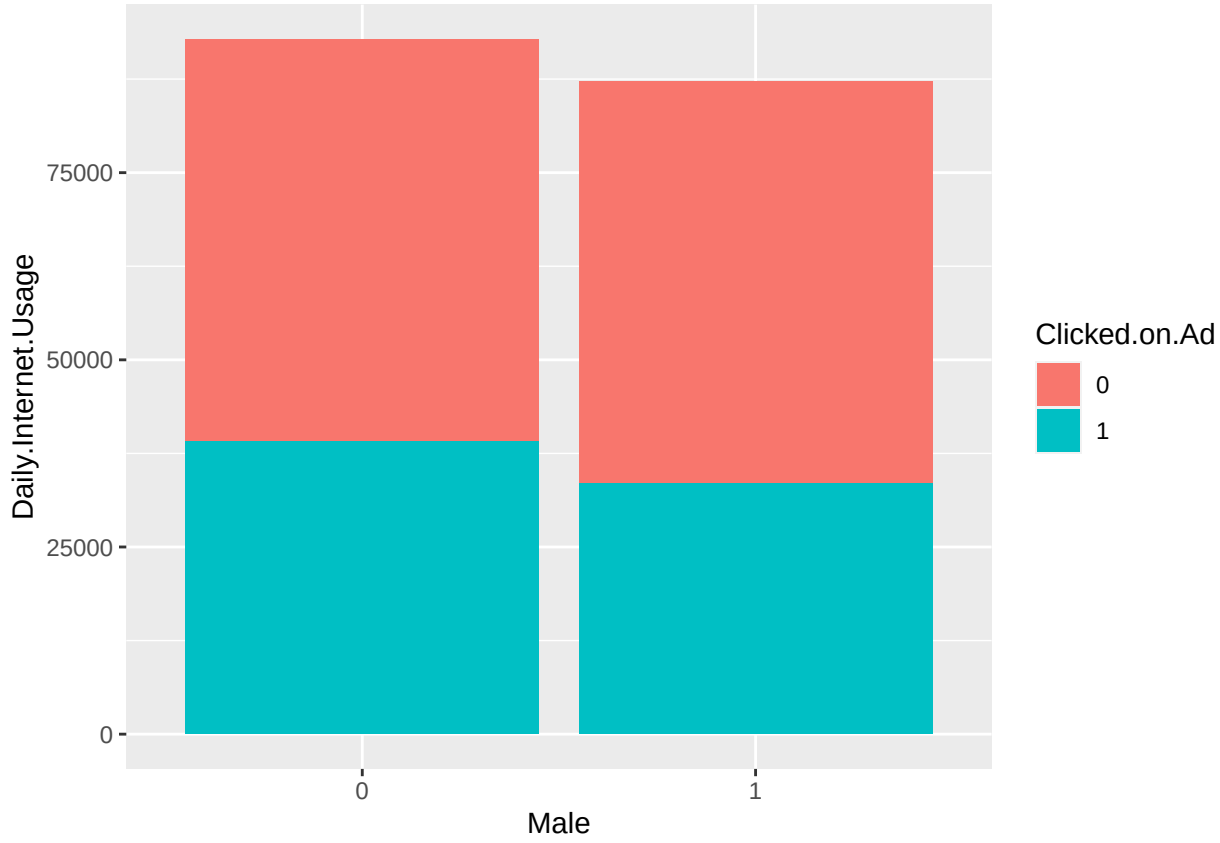



**Stacked column plots**

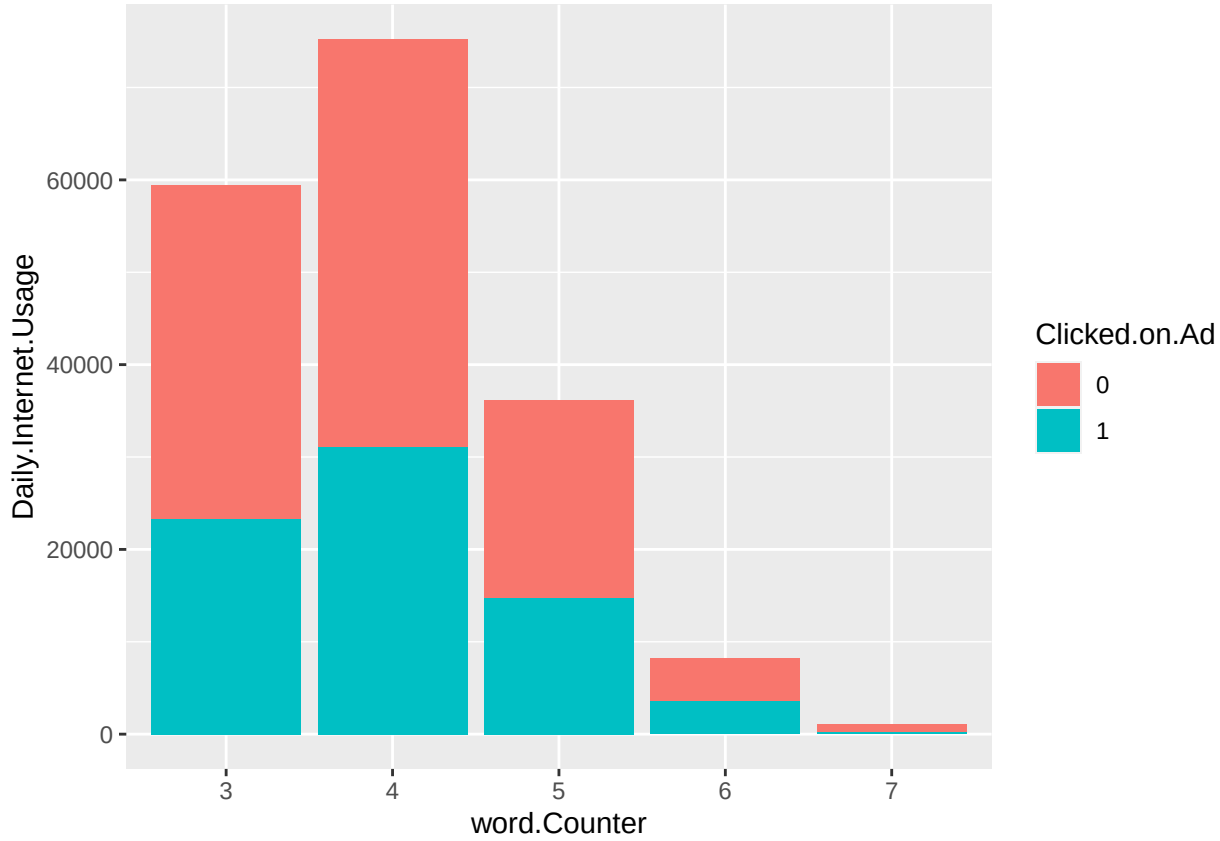
```
# Plotting Stacked column of male(gender) vs Area.Income
#
ggplot(ad_dataset, aes(fill=Clicked.on.Ad, y=Area.Income, x=Male)) +
  geom_bar(position='stack', stat='identity')
```




```
# Plotting Stacked column of male(gender) vs Daily.Internet.Usage
#
ggplot(ad_dataset, aes(fill=Clicked.on.Ad, y=Daily.Internet.Usage, x=Male)) +
  geom_bar(position='stack', stat='identity')
```

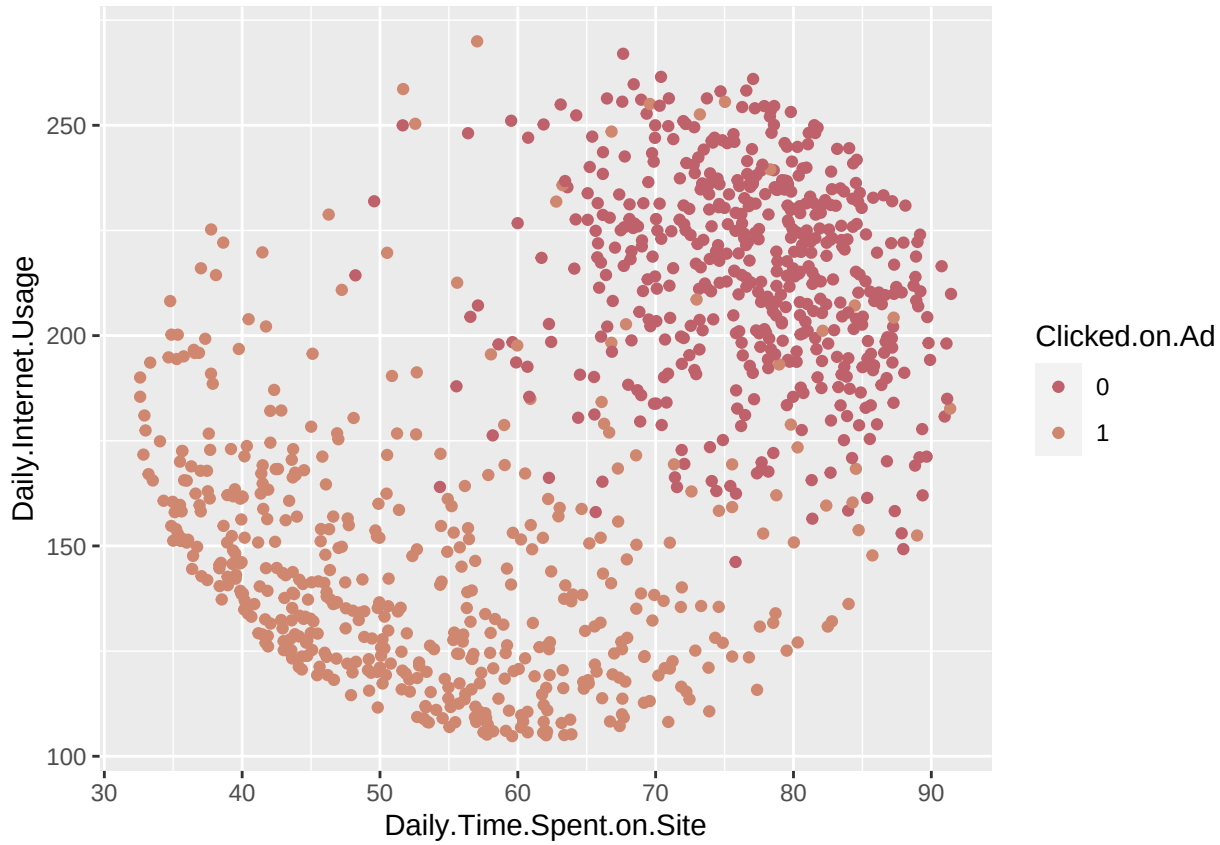



```
# Plotting Stacked column of word.Counter vs Daily.Internet.Usage
#
ggplot(ad_dataset, aes(fill=Clicked.on.Ad, y=Daily.Internet.Usage, x=word.Counter)) +
  geom_bar(position='stack', stat='identity')
```

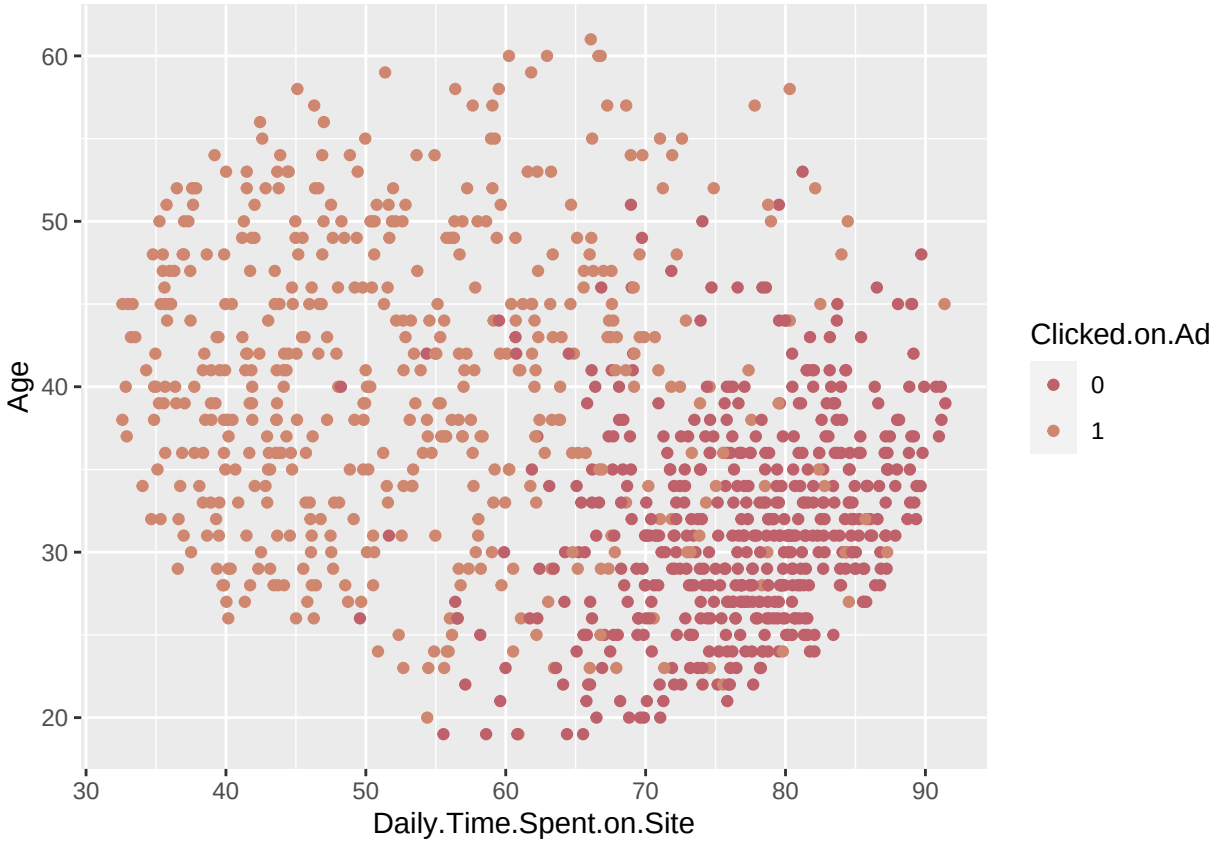



#### Bivariate Scatter Plots

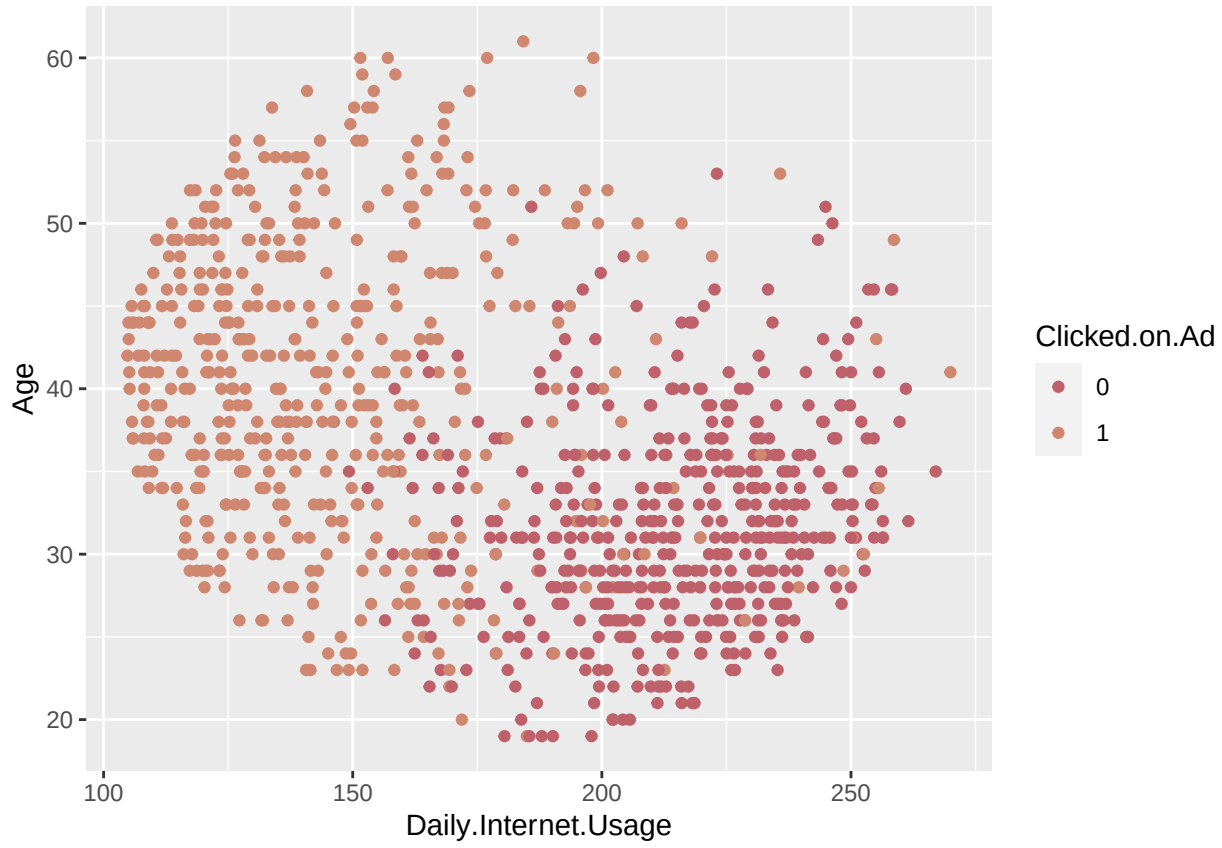
```
# Load the library palatteer
#
library(paletteer)

# Scatter plot of daily time spent on site vs daily internet usage
#
ggplot(ad_dataset, aes(Daily.Time.Spent.on.Site, Daily.Internet.Usage,
       color = Clicked.on.Ad)) + geom_point() + scale_color_paletteer_d("nord::aurora")
```




```
# scatter plot of daily time spent on the site vs age
#
ggplot(ad_dataset, aes(Daily.Time.Spent.on.Site, Age,
       color = Clicked.on.Ad)) + geom_point() + scale_color_paletteer_d("nord::aurora")
```

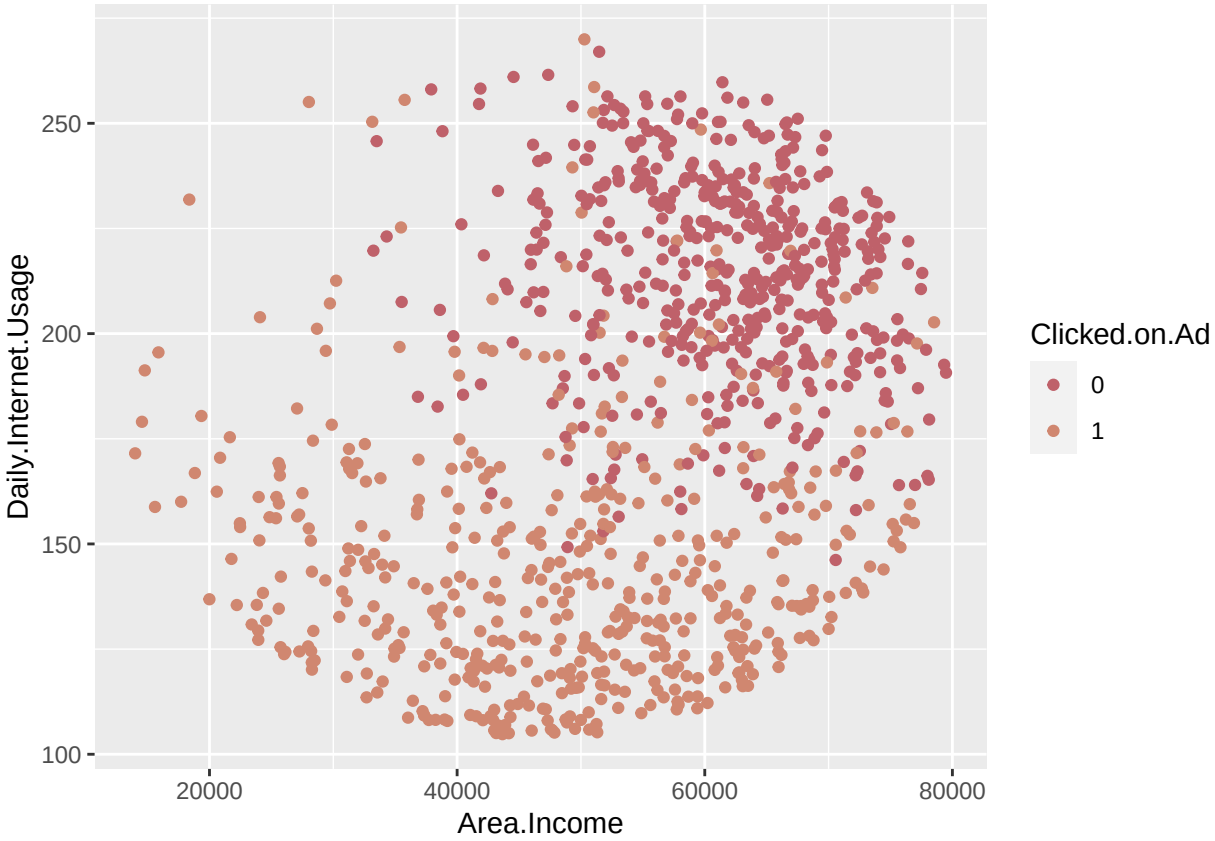



```
# Scatter plot of daily internet usage versus age
#
ggplot(ad_dataset, aes(Daily.Internet.Usage, Age,
       color = Clicked.on.Ad)) + geom_point() + scale_color_paletteer_d("nord::aurora")
```




```
# Scatter plot of area income vs age
#
ggplot(ad_dataset, aes(Area.Income, Age,
       color = Clicked.on.Ad)) + geom_point() + scale_color_paletteer_d("nord::aurora")
```


```
# scatter plot of daily internet usage vs area income
#
ggplot(ad_dataset, aes(Area.Income, Daily.Internet.Usage,
       color = Clicked.on.Ad)) + geom_point() + scale_color_paletteer_d("nord::aurora")
```




# Results ## Univariate analysis

From the univariate analysis;

Individuals that clicked the ads on average and mostly(mode) had lower area incomes, lower internet usage, higher ages, and lower times spent on site compared to those that did not click ads

The range, quantile, variance and standard deviation values show that the ages, daily time spent on site, daily internet usage, and area incomes vary more among individuals that click on the ad compared to those that do not

From the value counts obtained from gender, the number of males to those that did not click the add was almost similar to the female numbers observed.

The number of words making up each ad topic line were counted and the ad topic line dropped. Most of the ad topic lines were either 3, 4, 5 words long. There were ad topic lines 6 or 7 words long. Slightly more ad clicks were observed for ad topic lines 3, 4, or 5 words long than those that were not clicked. However, for ad topic lines 6 or 7 words long, more of these adds were not clicked relative to those that were clicked.

The countries were grouped into continents for easier analysis. There were a total of five continents;Africa, Americas, Asia, Europe, and Oceania. All continents aside from Asia recorded slightly higher ad clicks compared to those ads that were not clicked.

## Bivariate analysis results

From the bivariate analysis;

Age had a negative covariance and correlation to the other numerical variables in the data set. The other numerical variables had positive covariance and correlation amongst each other.

From the plots;

Gender did not affect ad clicks even in relation to other variables in the dataset.

An individual with small age and spent more time on the site is less was less likely to click the add

Individuals with high internet usage per day and spent more time on the site were observed not to click on the add on most occasions

Individuals with high daily internet usage and are young were observed to not click on ads mostly

Individuals who are young in areas with high income do not click on the ads most of the time

Individuals in areas with high income and have a high daily internet consumption do not on most occasions click on the ads.

# Conclusion

From the obtained results it can be concluded that;

Gender of the indivdual using the site does not affect the chances of the user clicking the ad. The ad is equal likely to be clicked or not clicked regardless whether the user is male or female.

Age, area income, daily internet usage, daily time spent on the site significantly affect whether one clicks on an add or not. Age however affects the ad clicks proportionally. The higher the age of the individual the higher the chances of him or her clicking the ad. Daily internet usage, time spent on the site and area income affect the chances of an ad being clicked on inversely. The higher either of these factors the lower the chances of this ad being clicked by the individual. This is evident in the inverse correlation between age and these three continuous variables

The continent, and the word length of thead topic line affect whether the ad is clicked or not but to a smaller extent. Asians are more likely not the click the ad compared to individuals from other continents. Ad topic lines longer than five words are most likely not to be clicked at all. Ad topic lines of 3 -5 words have slightly higher chances of being clicked rather than not being clicked.

# Reccomendations

From the analysis results and conclusion, the entrepreneur should target older individuals who come from areas with low area incomes, don't spend a lot of time on the site and have low daily internet usage.

The entrepreneur should avoid running ad topic lines with more than five words to minimize the chances of them not being clicked. She should also reduce the number of ads in the Asian market since their is a slightly higher chance of the not being clicked by an individual in Asia

In running her ads, gender should not factor into her decision making since from the analysis it has been affirmed that gender does not affect the chances of the ad being clicked or not.

# Further Questions

## A) Do we have the right data

For this study and to meet the objectives set by the entrepreneur, this data provides relevant information to meet those objectives.

## B) Do we have the right question?

For this problem, the question of determining the individuals most likely to click her ads is correct. Since by answering this question the entrepreneur can maximize the reach of her adverts.