R-Fundermentals Independent Project - Week 12

Mary Njuguna

2022-05-29

Research Question

To identify which individuals and major factors that will contribute to the clicking of the ads.

Metric Of susccess

To identify the factors that contribute successful ad click via performing univarite and bivariate analysis

Experimental Design

- Importing libraries
- Loading the dataset
- Cleaning the data by checking and dealing with missing values, duplicates, unique values and outliers.
- Performing univariate analysis
- Performing bivariate analysis
- Conclusion
- Recommendation

Data Appropriateness

The data was appropriate for the study.

```
# Importing libraries
library(ggvis)
#library(tidyverse)
#library(ggot2pl)
library(tibble)
```

```
 #Loading the data set

df<-read.csv("http://bit.ly/IPAdvertisingData")

# Converting the data set to be in tibble form. The advantages of using tibble is that, it never conver

df<-as_tibble(df)
# Previewing the first six rows
head(df)
```

```
## # A tibble: 6 x 10
##   Daily.Time.Spent~   Age Area.Income Daily.Internet.~ Ad.Topic.Line City   Male
##               <dbl> <int>       <dbl>            <dbl> <chr>         <chr> <int>
## 1              69.0    35      61834.             256. Cloned 5thge~ Wrig~     0
## 2              80.2    31      68442.             194. Monitored na~ West~     1
## 3              69.5    26      59786.             236. Organic bott~ Davi~     0
## 4              74.2    29      54806.             246. Triple-buffe~ West~     1
## 5              68.4    35      73890.             226. Robust logis~ Sout~     0
## 6              60.0    23      59762.             227. Sharable cli~ Jami~     1
## # ... with 3 more variables: Country <chr>, Timestamp <chr>,
## #   Clicked.on.Ad <int>
```

```
#Exploring the bottom of the dataset
tail(df)
```

```
## # A tibble: 6 x 10
##   Daily.Time.Spent~   Age Area.Income Daily.Internet.~ Ad.Topic.Line City   Male
##               <dbl> <int>       <dbl>            <dbl> <chr>         <chr> <int>
## 1              43.7    28      63127.             173. Front-line b~ Nich~     0
## 2              73.0    30      71385.             209. Fundamental ~ Duff~     1
## 3              51.3    45      67782.             134. Grass-roots ~ New ~     1
## 4              51.6    51      42416.             120. Expanded int~ Sout~     1
## 5              55.6    19      41921.             188. Proactive ba~ West~     0
## 6              45.0    26      29876.             178. Virtual 5thg~ Ronn~     0
## # ... with 3 more variables: Country <chr>, Timestamp <chr>,
## #   Clicked.on.Ad <int>
```

```
#Determining the class of the dataset
class(df)
```

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

```
#Determining the shape of the dataset
dim(df)
```

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

```
#Determining the structure of the dataset
str(df)
```

```
## tibble [1,000 x 10] (S3: tbl_df/tbl/data.frame)
##  $ Daily.Time.Spent.on.Site: num [1:1000] 69 80.2 69.5 74.2 68.4 ...
##  $ Age                     : int [1:1000] 35 31 26 29 35 23 33 48 30 20 ...
##  $ Area.Income             : num [1:1000] 61834 68442 59786 54806 73890 ...
##  $ Daily.Internet.Usage    : num [1:1000] 256 194 236 246 226 ...
##  $ Ad.Topic.Line           : chr [1:1000] "Cloned 5thgeneration orchestration" "Monitored national st
##  $ City                    : chr [1:1000] "Wrightburgh" "West Jodi" "Davidton" "West Terrifurt" ...
##  $ Male                    : int [1:1000] 0 1 0 1 0 1 0 1 1 1 ...
##  $ Country                 : chr [1:1000] "Tunisia" "Nauru" "San Marino" "Italy" ...
##  $ Timestamp               : chr [1:1000] "2016-03-27 00:53:11" "2016-04-04 01:39:02" "2016-03-13 20
##  $ Clicked.on.Ad           : int [1:1000] 0 0 0 0 0 0 0 1 0 0 ...
```

```
#Listing the columns of the dataset
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"
```

Cleaning The Dataset

```
#Checking for total missing values
sum(is.na(df))
```

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

The dataset does not have null values

```
#Checking for the duplicates
anyDuplicated(df)
```

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

The dataset does not have duplicates

```
#Extracting numeric values unorder to determine unique values and outliers
numeric_values<- unlist(lapply(df, is.numeric))

numeric_cols<-df[,numeric_values]
#Printing the numeric columns
head(numeric_cols)
```

```
## # A tibble: 6 x 6
##   Daily.Time.Spent.on.Si~   Age Area.Income Daily.Internet.~  Male Clicked.on.Ad
##                     <dbl> <int>       <dbl>            <dbl> <int>         <int>
## 1                    69.0    35      61834.             256.     0             0
## 2                    80.2    31      68442.             194.     1             0
```

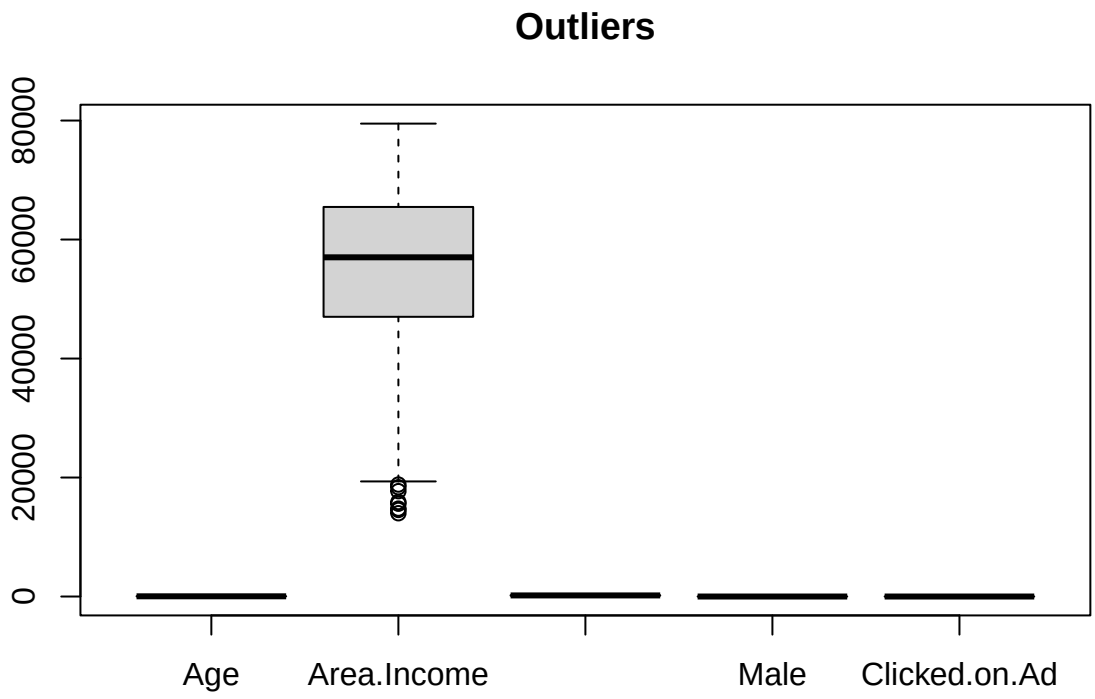
```
## 3                   69.5    26      59786.             236.     0             0
## 4                   74.2    29      54806.             246.     1             0
## 5                   68.4    35      73890.             226.     0             0
## 6                   60.0    23      59762.             227.     1             0
```

The tibble contains six numeric columns

```
#Checking for unique values
unique_values<- unique(numeric_cols)
head(numeric_cols)
```

```
## # A tibble: 6 x 6
##   Daily.Time.Spent.on.Si~   Age Area.Income Daily.Internet.~  Male Clicked.on.Ad
##                     <dbl> <int>       <dbl>            <dbl> <int>         <int>
## 1                    69.0    35      61834.             256.     0             0
## 2                    80.2    31      68442.             194.     1             0
## 3                    69.5    26      59786.             236.     0             0
## 4                    74.2    29      54806.             246.     1             0
## 5                    68.4    35      73890.             226.     0             0
## 6                    60.0    23      59762.             227.     1             0
```

```
#Plotting boxplot in order to visualize outliers
boxplot(numeric_cols[,-1],main="Outliers")
```



Income has outliers

```
#Checking the actual outliers
```

EDA

Univariate Analysis

###Getting the summary statistic of the numeric columns

```
summary(numeric_cols)
```

```
##  Daily.Time.Spent.on.Site      Age         Area.Income    Daily.Internet.Usage
##  Min.   :32.60            Min.   :19.00   Min.   :13996   Min.   :104.8
##  1st Qu.:51.36            1st Qu.:29.00   1st Qu.:47032   1st Qu.:138.8
##  Median :68.22            Median :35.00   Median :57012   Median :183.1
##  Mean   :65.00            Mean   :36.01   Mean   :55000   Mean   :180.0
##  3rd Qu.:78.55            3rd Qu.:42.00   3rd Qu.:65471   3rd Qu.:218.8
##  Max.   :91.43            Max.   :61.00   Max.   :79485   Max.   :270.0
##       Male       Clicked.on.Ad
##  Min.   :0.000   Min.   :0.0
##  1st Qu.:0.000   1st Qu.:0.0
##  Median :0.000   Median :0.5
##  Mean   :0.481   Mean   :0.5
##  3rd Qu.:1.000   3rd Qu.:1.0
##  Max.   :1.000   Max.   :1.0
```

- The mean of age is 36.01, the youngest being 19years old and the oldest being 61 years old.
- the maxmum daily time spent on site is 91.43, the minimum being 32.60 and the mean is 65.
- The maxmum daily internet used is 270, the minimum is 104.8 and the men is 180.
- The mean of the clicked add is 0.5
- The maxmum income of the individuals is 79485, the minimum is 13996 and the mean is 55000

Mode in the numeric colums

```
#Value that appeared most frequently in daily.time.spent.on.site
getmode<-function(v){
  uniqv<-unique(v)
  uniqv[which.max(tabulate(match(v,uniqv)))]
}
time.mode<-getmode(numeric_cols$Daily.Time.Spent.on.Site)
time.mode
```

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

- The most time which was spent on the page was 62.26

```r
#Value that appeared most frequently in age
getmode<-function(v){
  uniqv<-unique(v)
  uniqv[which.max(tabulate(match(v,uniqv)))]
}
age.mode<-getmode(numeric_cols$Age)
age.mode
```

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

- Most of the people who clicked the page were 31 years of age.

```r
#Value that appeared most frequently in daily.internet.usage
getmode<-function(v){
  uniqv<-unique(v)
  uniqv[which.max(tabulate(match(v,uniqv)))]
}
internet.mode<-getmode(numeric_cols$Daily.Internet.Usage)
internet.mode
```

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

- Most of the daily internet use was 167.22

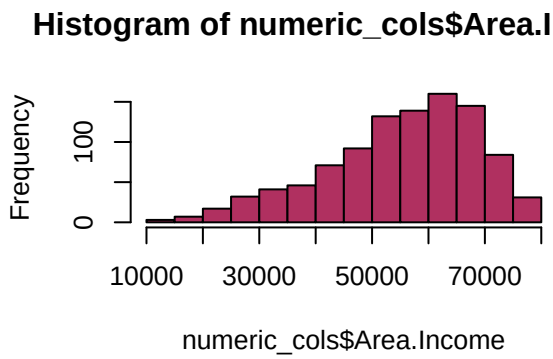
```r
#Value that appeared most frequently in area.income
getmode<-function(v){
  uniqv<-unique(v)
  uniqv[which.max(tabulate(match(v,uniqv)))]
}
income.mode<-getmode(numeric_cols$Area.Income)
income.mode
```

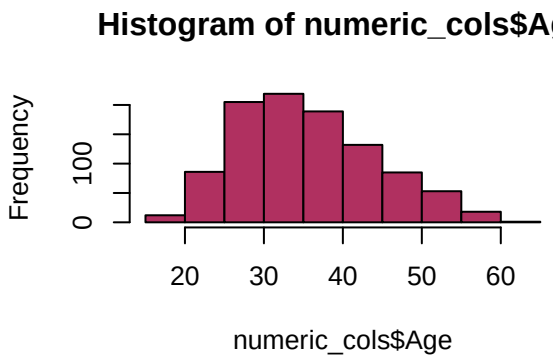
```
## [1] 61833.9
```

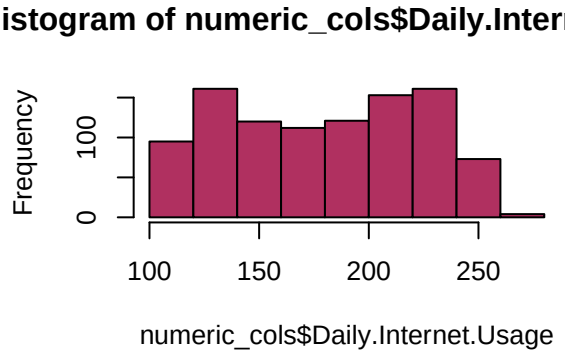
Plotting of the graphs

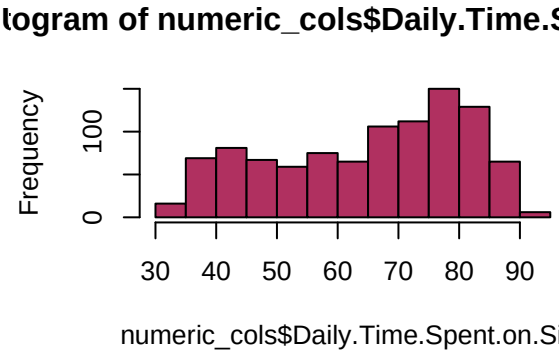
Bar plot of the numeric values

```r
par(mfrow = c(2,2))
hist(numeric_cols$Area.Income, col = "Maroon")
hist(numeric_cols$Age,col = "Maroon")
hist(numeric_cols$Daily.Internet.Usage,col = "Maroon")
hist(numeric_cols$Daily.Time.Spent.on.Site,col = "Maroon")
```









- most individuals has income ranging between 50000 and 70000 - The daily internet is almost uniform - most of the people who visits the site are between 30 to 40 years.

```
colnames(numeric_cols)
```

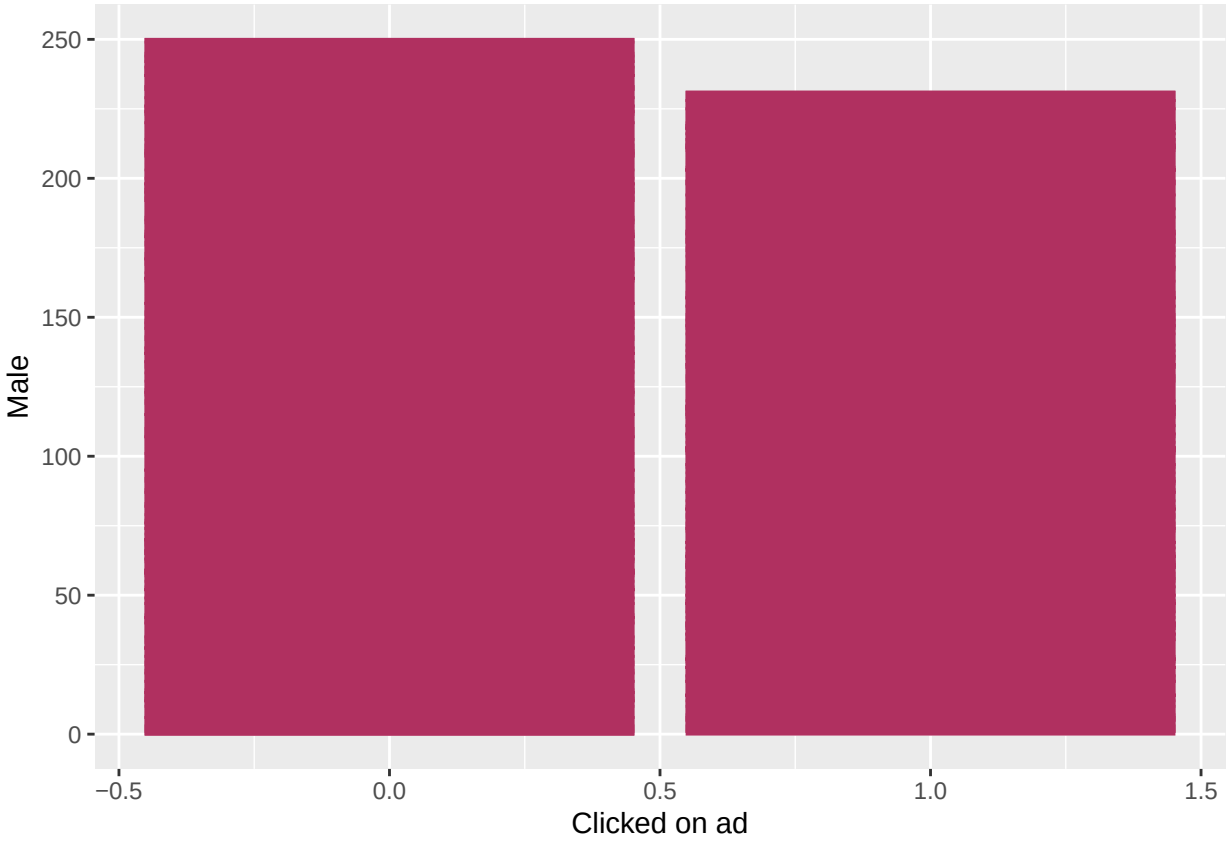
```
## [1] "Daily.Time.Spent.on.Site" "Age"
## [3] "Area.Income"              "Daily.Internet.Usage"
## [5] "Male"                     "Clicked.on.Ad"
```

```
#Bar plot between clicked ad and Male
library(ggplot2)
```

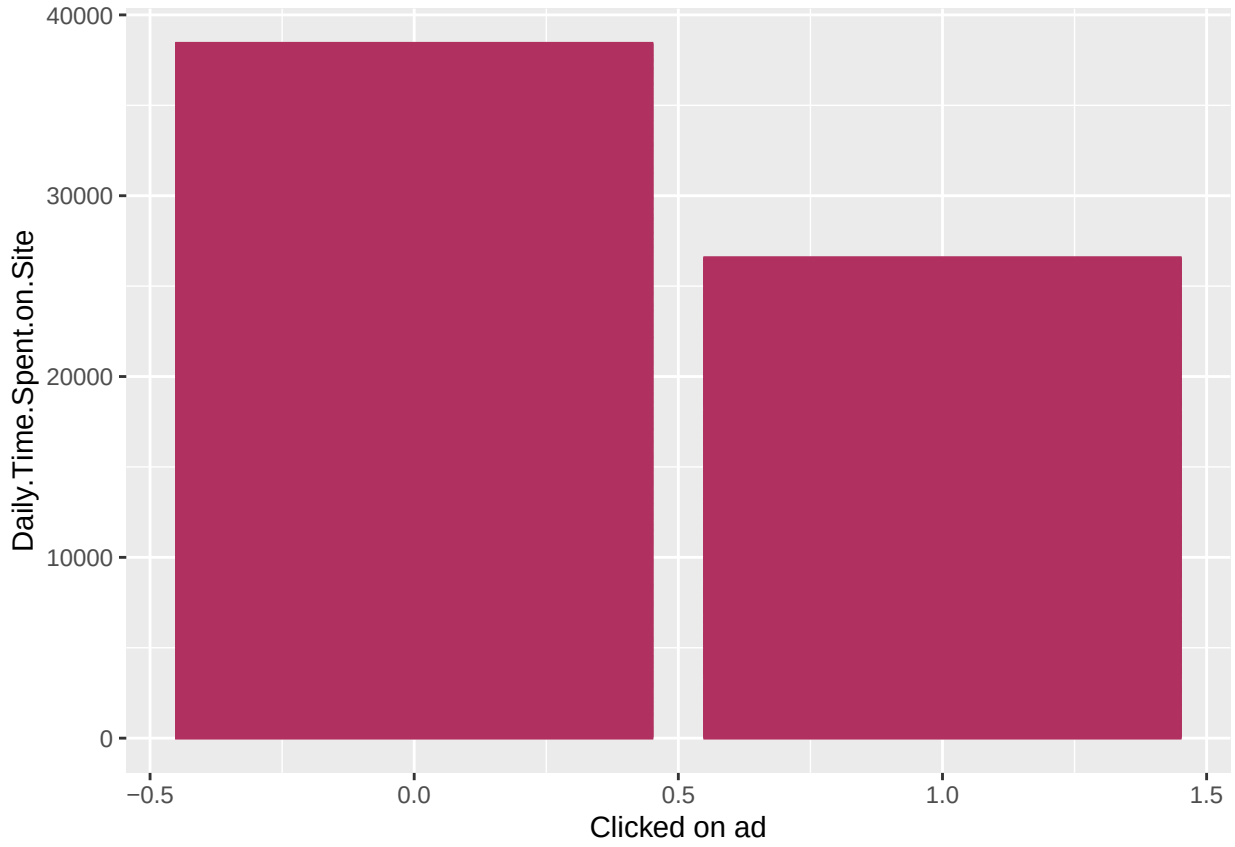
```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:ggvis':
##
##     resolution
```

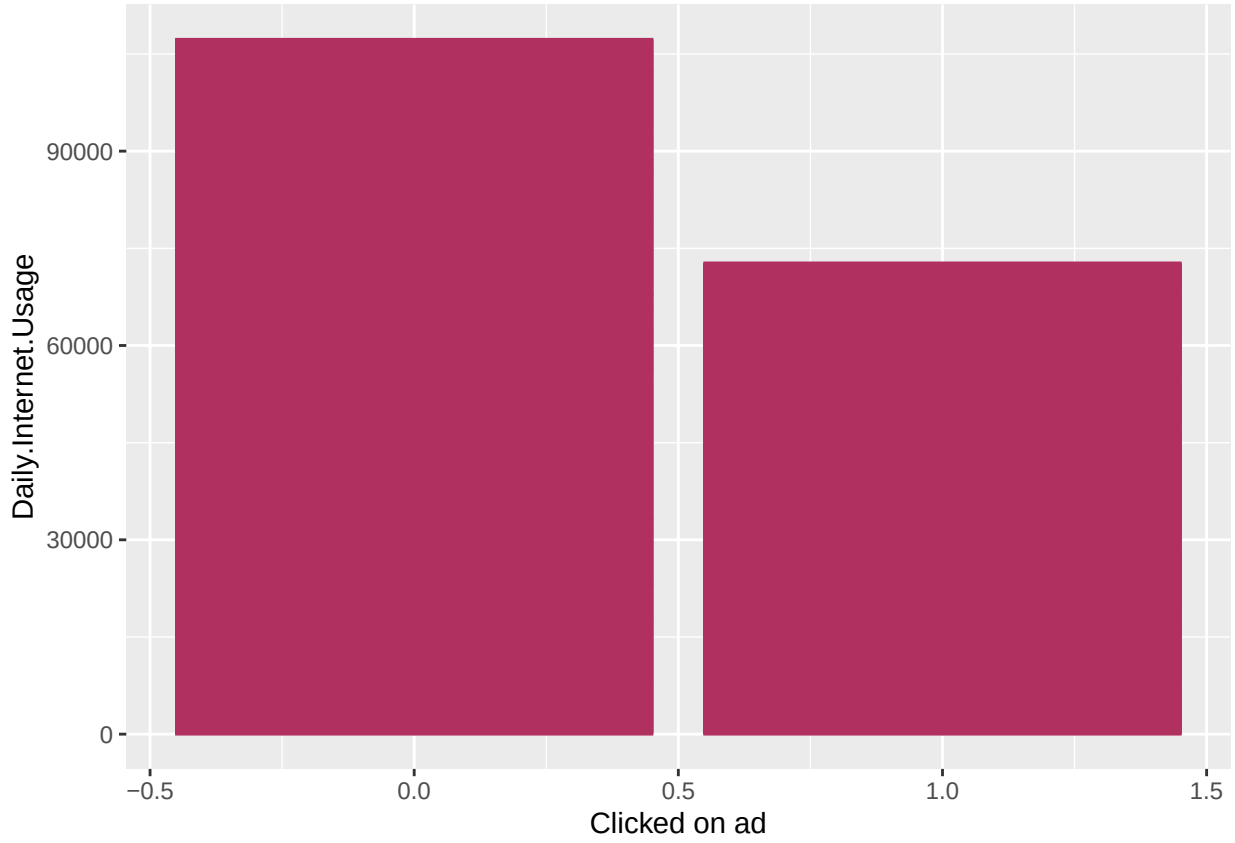
```
ggplot(numeric_cols, aes(Clicked.on.Ad,Male )) +
   geom_bar(stat = 'identity',col = "Maroon")+ labs(y="Male",x="Clicked on ad")
```



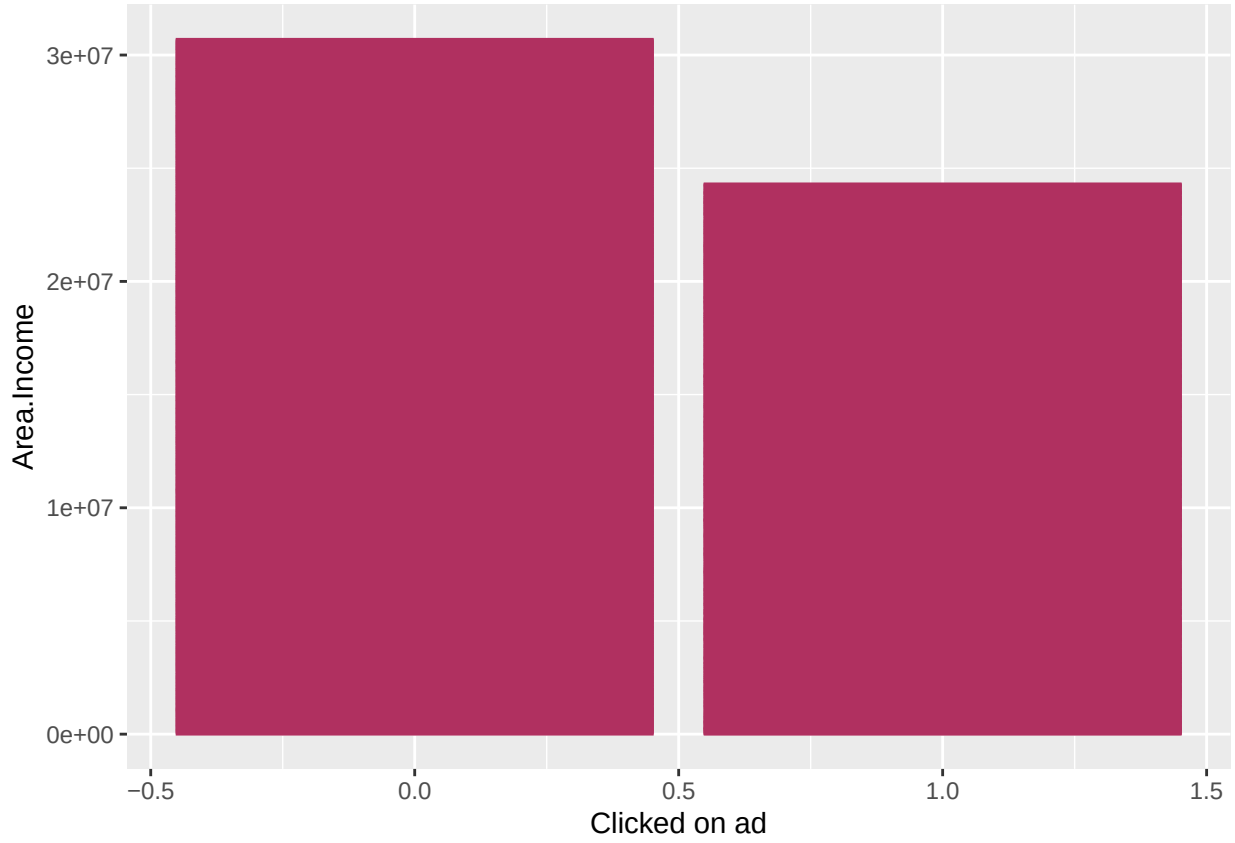

```
#Bar plot between clicked ad andDaily.Time.Spent.on.Site
ggplot(numeric_cols, aes(Clicked.on.Ad,Daily.Time.Spent.on.Site )) +
   geom_bar(stat = 'identity',col="Maroon")+ labs(y="Daily.Time.Spent.on.Site",x="Clicked on ad")
```



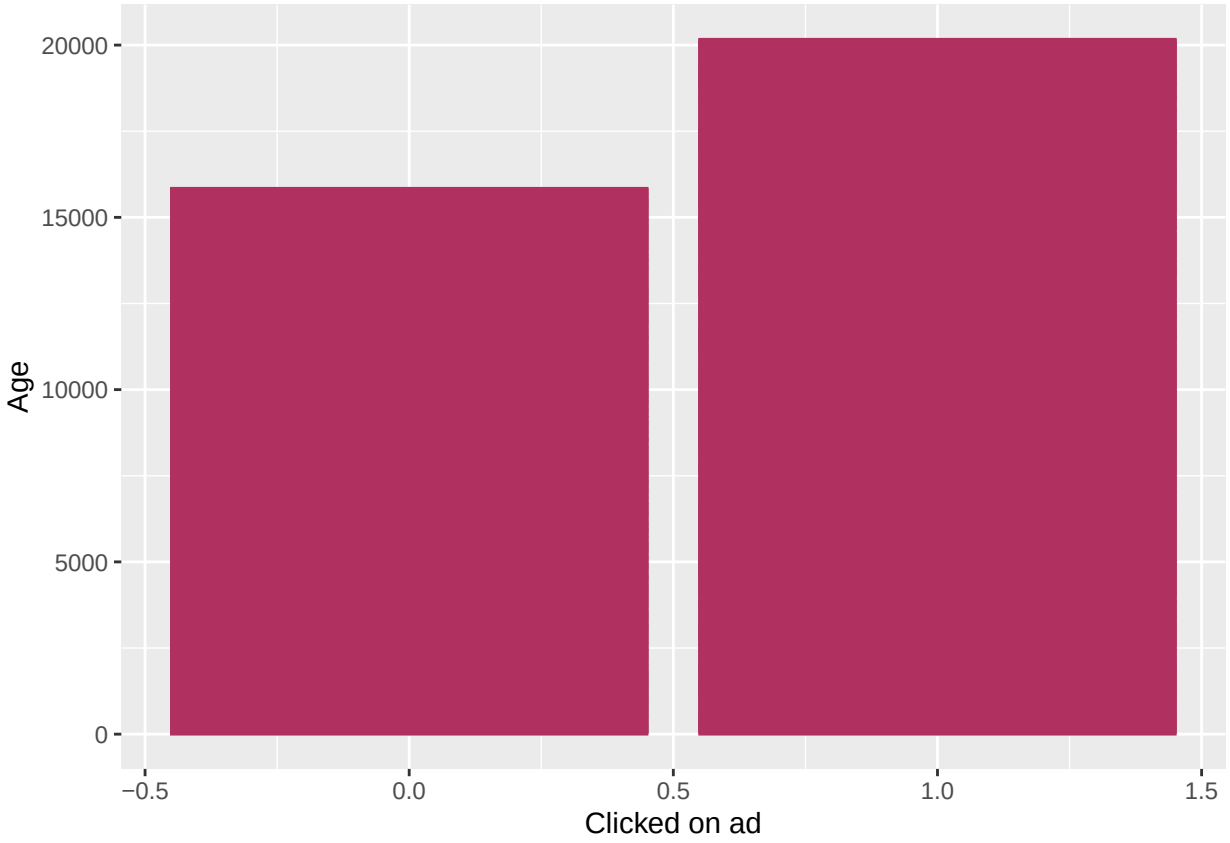

```
#Bar plot between clicked ad and Daily.Internet.Usage
ggplot(numeric_cols, aes(Clicked.on.Ad,Daily.Internet.Usage)) +
   geom_bar(stat = 'identity',col="Maroon")+ labs(y="Daily.Internet.Usage",x="Clicked on ad")
```




```
#Bar plot between clicked ad and Area.Income
ggplot(numeric_cols, aes(Clicked.on.Ad,Area.Income)) +
   geom_bar(stat = 'identity',col="Maroon")+ labs(y="Area.Income",x="Clicked on ad")
```




```
#Bar plot between clicked ad and Age
ggplot(numeric_cols, aes(Clicked.on.Ad,Age)) +
   geom_bar(stat = 'identity',col="Maroon")+ labs(y="Age",x="Clicked on ad")
```



Covariance

```
# Assigning the Daily.Time.Spent.on.Site column to the variable dailt_time
daily_time <- numeric_cols$Daily.Time.Spent.on.Site
# Assigning the Age column to the variable waiting
age<- numeric_cols$Age
# Using the cov() function to determine the age
cov(daily_time, age)
```

```
## [1] -46.17415
```

the covariance relationship between Daily time spent on site and Age is -46.17415 indicating a negative linear relationship between the two variables

```
# Assigning the Clicked.on.Ad column to the variable Ad_clicked
Ad_clicked <- numeric_cols$Clicked.on.Ad
# Assigning the Age column to the variable age
age<- numeric_cols$Age
# Using the cov() function to determine the covariance
cov(Ad_clicked, age)
```

```
## [1] 2.164665
```

The covariance relationship between Ad clicked and Age is 2.164665 indicating weak positive linear relationship between the two variables

```
# Assigning the Clicked.on.Ad column to the variable Ad_clicked
Ad_clicked <- numeric_cols$Clicked.on.Ad
# Assigning the Daily.Time.Spent.on.Site column to the variable daily_time
daily_time<- numeric_cols$Daily.Time.Spent.on.Site
# Using the cov() function to determine the covariance
cov(Ad_clicked, daily_time)
```

```
## [1] -5.933143
```

The covariance relationship between Ad_clicked and daily_time spent on site is -5.933143 indicating negative linear relationship between the two variables.

```
# Assigning the Clicked.on.Ad column to the variable Ad_clicked
Ad_clicked <- numeric_cols$Clicked.on.Ad
# Assigning the Daily.Internet.Usage column to the variable internet_usage
internet_usage<- numeric_cols$Daily.Internet.Usage
# Using the cov() function to determine the covariance
cov(Ad_clicked, internet_usage)
```

```
## [1] -17.27409
```

The covariance relationship between Ad_clicked and Daily internet usage is -17.27409 indicating negative linear relationship between the two variables.

Correlation coefficient

```
# Assigning the Daily.Time.Spent.on.Site column to the variable dailt_time
daily_time <- numeric_cols$Daily.Time.Spent.on.Site
# Assigning the Age column to the variable waiting
age<- numeric_cols$Age
# Using the cov() function to determine the age
cor(daily_time, age)
```

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

the correlation relationship between Daily time spent on site and Age is -0.3315133 indicating a negative linear relationship between the two variables

```
# Assigning the Clicked.on.Ad column to the variable Ad_clicked
Ad_clicked <- numeric_cols$Clicked.on.Ad
# Assigning the Age column to the variable age
age<- numeric_cols$Age
# Using the cov() function to determine the covariance
cor(Ad_clicked, age)
```

```
## [1] 0.4925313
```

The covariance relationship between Ad clicked and Age is 0.4925313 indicating weak positive linear relationship between the two variables.

```
# Assigning the Clicked.on.Ad column to the variable Ad_clicked
Ad_clicked <- numeric_cols$Clicked.on.Ad
# Assigning the Daily.Time.Spent.on.Site column to the variable daily_time
daily_time<- numeric_cols$Daily.Time.Spent.on.Site
# Using the cov() function to determine the covariance
cor(Ad_clicked, daily_time)
```

```
## [1] -0.7481166
```

The correlation between Ad_clicked and daily_time spent on site is -0.7481166 indicating negative linear relationship between the two variables.

```
# Assigning the Clicked.on.Ad column to the variable Ad_clicked
Ad_clicked <- numeric_cols$Clicked.on.Ad
# Assigning the Daily.Internet.Usage column to the variable internet_usage
internet_usage<- numeric_cols$Daily.Internet.Usage
# Using the cov() function to determine the covariance
cor(Ad_clicked, internet_usage)
```

```
## [1] -0.7865392
```

The correlation between Ad_clicked and Daily internet usage is -0.7865392 indicating negative linear relationship between the two variables.

Summary of correlation

```
#Summary of correlation
numeric.corr = cor(numeric_cols)
numeric.corr
```

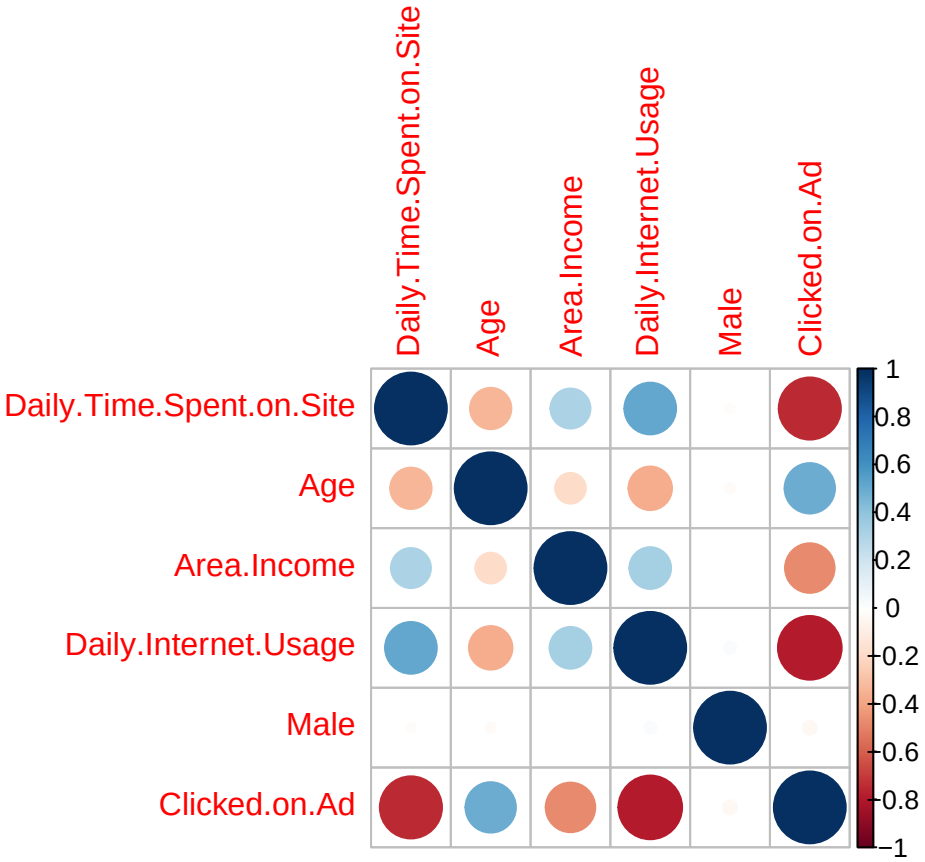
```
##                           Daily.Time.Spent.on.Site         Age  Area.Income
## Daily.Time.Spent.on.Site               1.00000000 -0.33151334  0.310954413
## Age                                   -0.33151334  1.00000000 -0.182604955
## Area.Income                            0.31095441 -0.18260496  1.000000000
## Daily.Internet.Usage                   0.51865848 -0.36720856  0.337495533
## Male                                  -0.01895085 -0.02104406  0.001322359
## Clicked.on.Ad                         -0.74811656  0.49253127 -0.476254628
##                           Daily.Internet.Usage         Male Clicked.on.Ad
## Daily.Time.Spent.on.Site           0.51865848 -0.018950855   -0.74811656
## Age                               -0.36720856 -0.021044064    0.49253127
## Area.Income                        0.33749553  0.001322359   -0.47625463
## Daily.Internet.Usage               1.00000000  0.028012326   -0.78653918
## Male                               0.02801233  1.000000000   -0.03802747
## Clicked.on.Ad                     -0.78653918 -0.038027466    1.00000000
```

Visualizing the correlation matrix

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(numeric.corr)
```



Daily internet use and daily time spent on site has higher correlation. Also clicked ad and also age has higher correlation.

```
colnames(numeric_cols)
```

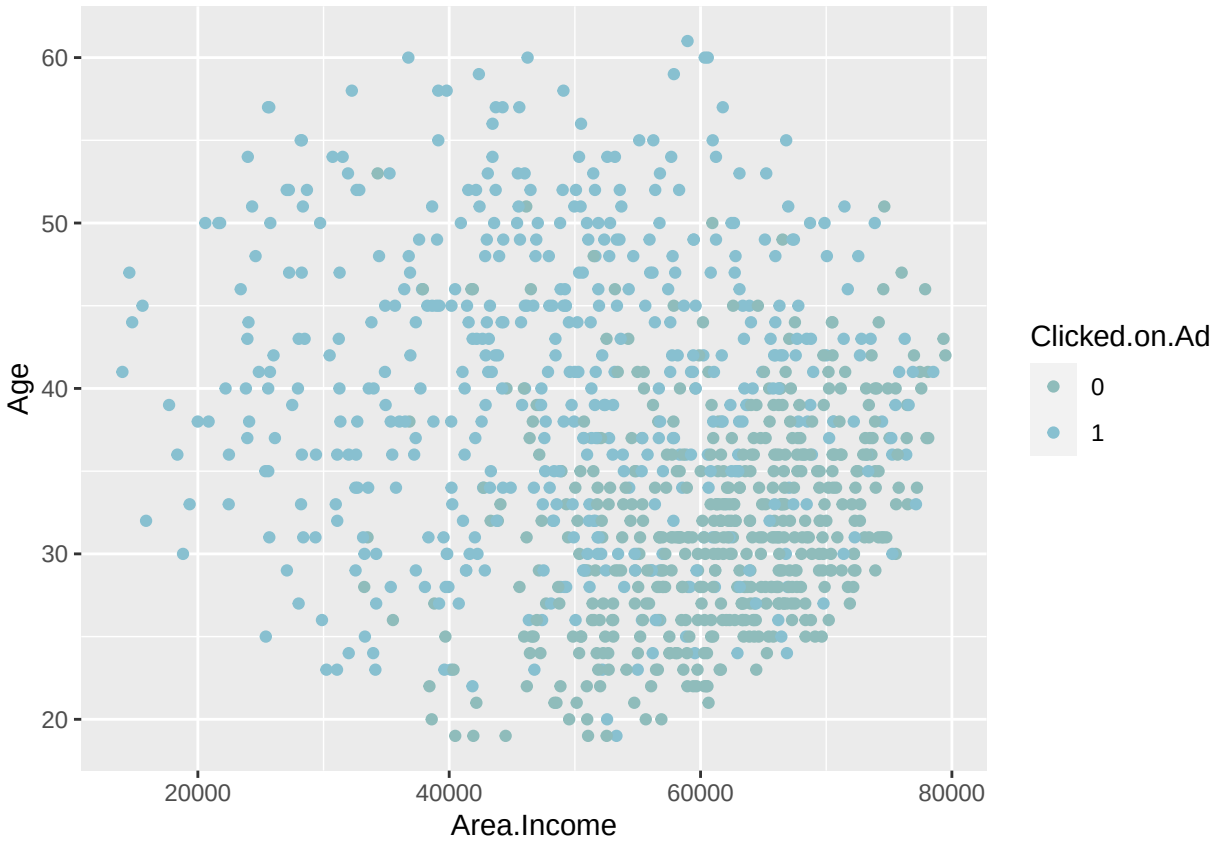
```
## [1] "Daily.Time.Spent.on.Site" "Age"
## [3] "Area.Income"              "Daily.Internet.Usage"
## [5] "Male"                     "Clicked.on.Ad"
```

converting target variable into factor for plotting

```
as.factor(numeric_cols$Clicked.on.Ad)->numeric_cols$Clicked.on.Ad
```

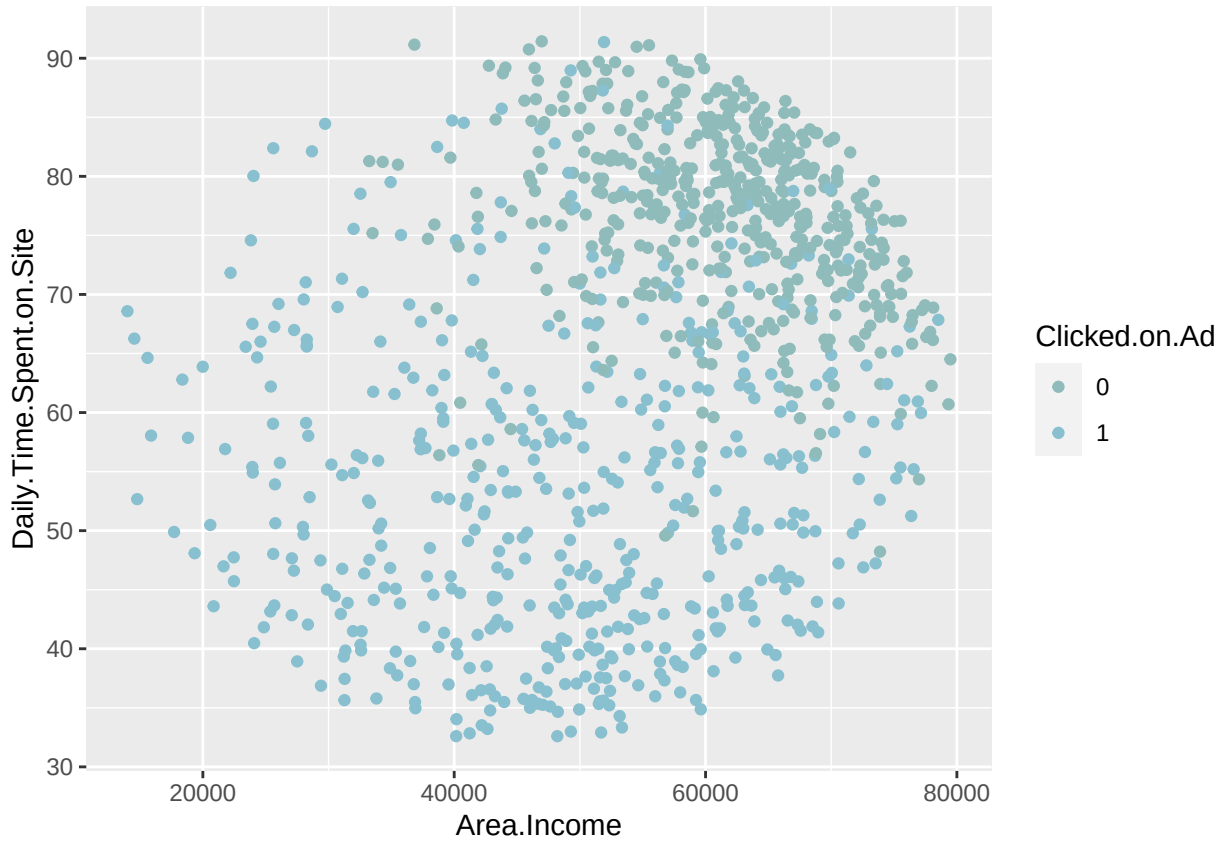
Scatter plots

```
#Scatter plot of area income and age
library(ggplot2)
library(paletteer)
ggplot(numeric_cols,aes(Area.Income,Age,
                        color =Clicked.on.Ad)) +geom_point()+scale_color_paletteer_d("nord::frost")
```

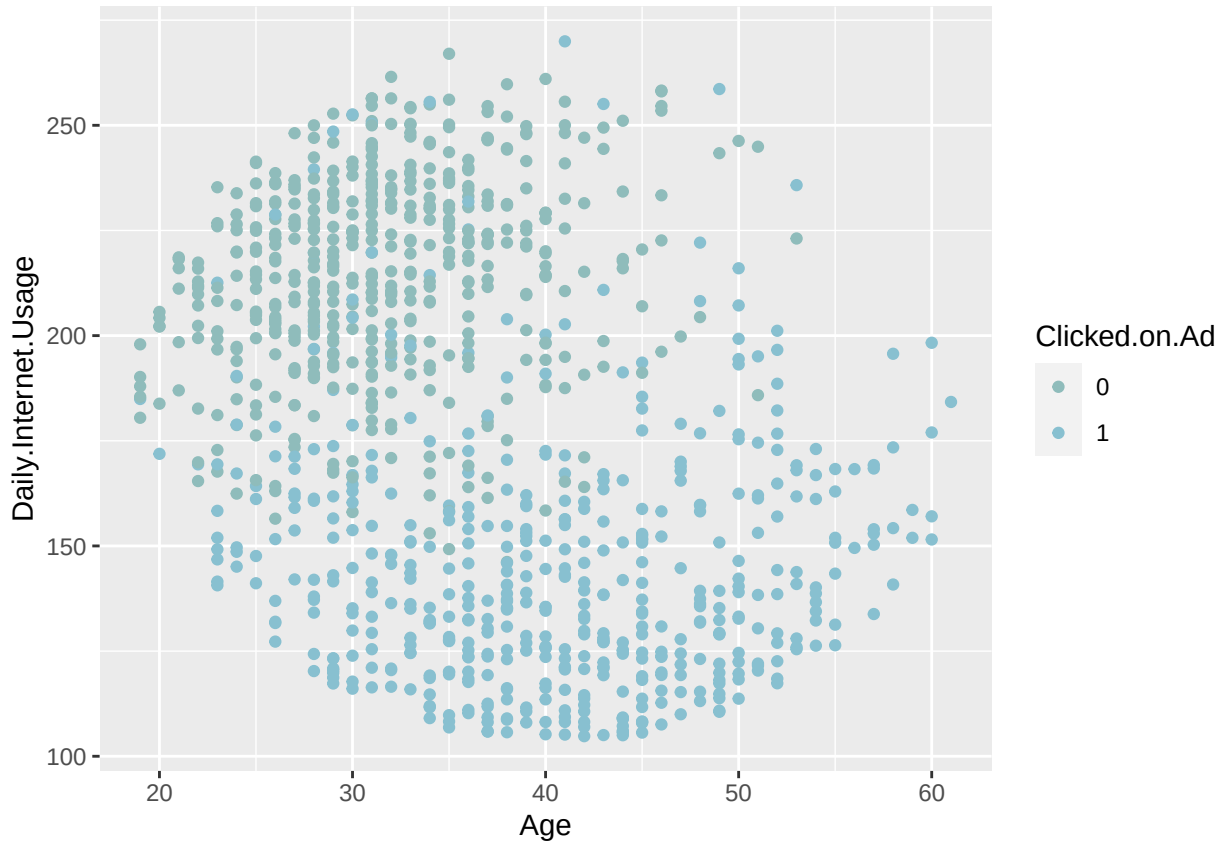


Age does note contribute much with the number of ads clicked. the larger the income the more ads are clicked.

```
#Scatter plot of area income and age
library(ggplot2)
library(paletteer)
ggplot(numeric_cols,aes(Area.Income,Daily.Time.Spent.on.Site,
                        color =Clicked.on.Ad)) +geom_point()+scale_color_paletteer_d("nord::frost")
```




```
#Scatter plot of area income and age
library(ggplot2)
library(paletteer)
ggplot(numeric_cols,aes(Age,Daily.Internet.Usage,
                        color =Clicked.on.Ad)) +geom_point()+scale_color_paletteer_d("nord::frost")
```



Ectracting categorical variables

```
#Printing the numeric columns


category<-unlist(lapply(df, is.character))

category_cols<-df[,category]
head(category_cols)
```

```
## # A tibble: 6 x 4
##   Ad.Topic.Line                          City           Country    Timestamp
##   <chr>                                  <chr>          <chr>      <chr>
## 1 Cloned 5thgeneration orchestration     Wrightburgh    Tunisia    2016-03-27 00~
## 2 Monitored national standardization     West Jodi      Nauru      2016-04-04 01~
## 3 Organic bottom-line service-desk       Davidton       San Marino 2016-03-13 20~
## 4 Triple-buffered reciprocal time-frame  West Terrifurt  Italy      2016-01-10 02~
## 5 Robust logistical utilization          South Manuel   Iceland    2016-06-03 03~
## 6 Sharable client-driven software        Jamieberg      Norway     2016-05-19 14~
```

```
#importing library
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
```

```
#Extracting categorical values
category = df %>% select_if(is.character)

#printing the categorical values
head(category)
```

```
## # A tibble: 6 x 4
##   Ad.Topic.Line                          City           Country    Timestamp
##   <chr>                                  <chr>          <chr>      <chr>
## 1 Cloned 5thgeneration orchestration     Wrightburgh    Tunisia    2016-03-27 00~
## 2 Monitored national standardization     West Jodi      Nauru      2016-04-04 01~
## 3 Organic bottom-line service-desk       Davidton       San Marino 2016-03-13 20~
## 4 Triple-buffered reciprocal time-frame  West Terrifurt  Italy      2016-01-10 02~
## 5 Robust logistical utilization          South Manuel   Iceland    2016-06-03 03~
## 6 Sharable client-driven software        Jamieberg      Norway     2016-05-19 14~
```

```
#Listing the columns which are categorical
colnames(category)
```

```
## [1] "Ad.Topic.Line" "City"          "Country"       "Timestamp"
```

```
country <- category$Country

country_frequency <-table(country)
s<-desc(country_frequency)
head(s,n=3)
```

```
## country
## Afghanistan     Albania     Algeria
##          -8          -7          -6
```

Conclusion

- the daily time spent on site was not directly proportional to the number of Ads clicked
- Age determined a lot the number of clicked ads. Aged people clicked more Ads
- Daily internet used was not directly proportional with the number of clicked Ads

Recommendation

- The types of Ads should be made relevasnce to all age groups in order to attract more people to click them.