#Define the Question a) Specify the Question Get insights on the data

b)Metric of success Determining correctly who is most likely to click on the ad

c) 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.

d)Experimental Design I shall incorporate the CRISPS-DM in my analysis

```
#upoad the dataset
setwd("C:/Users/Maureen M/Desktop/advert")
df = read.csv("advertising.csv.csv")
#read the dataset
head(df)
```

```
##   Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage
## 1                    68.95  35    61833.90               256.09
## 2                    80.23  31    68441.85               193.77
## 3                    69.47  26    59785.94               236.50
## 4                    74.15  29    54806.18               245.89
## 5                    68.37  35    73889.99               225.58
## 6                    59.99  23    59761.56               226.74
##                           Ad.Topic.Line           City Male    Country
## 1    Cloned 5thgeneration orchestration    Wrightburgh    0    Tunisia
## 2    Monitored national standardization      West Jodi    1      Nauru
## 3      Organic bottom-line service-desk       Davidton    0 San Marino
## 4 Triple-buffered reciprocal time-frame West Terrifurt    1      Italy
## 5         Robust logistical utilization   South Manuel    0    Iceland
## 6       Sharable client-driven software      Jamieberg    1     Norway
##             Timestamp Clicked.on.Ad
## 1 2016-03-27 00:53:11             0
## 2 2016-04-04 01:39:02             0
## 3 2016-03-13 20:35:42             0
## 4 2016-01-10 02:31:19             0
## 5 2016-06-03 03:36:18             0
## 6 2016-05-19 14:30:17             0
```

```
tail(df)
```

```
##      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
```

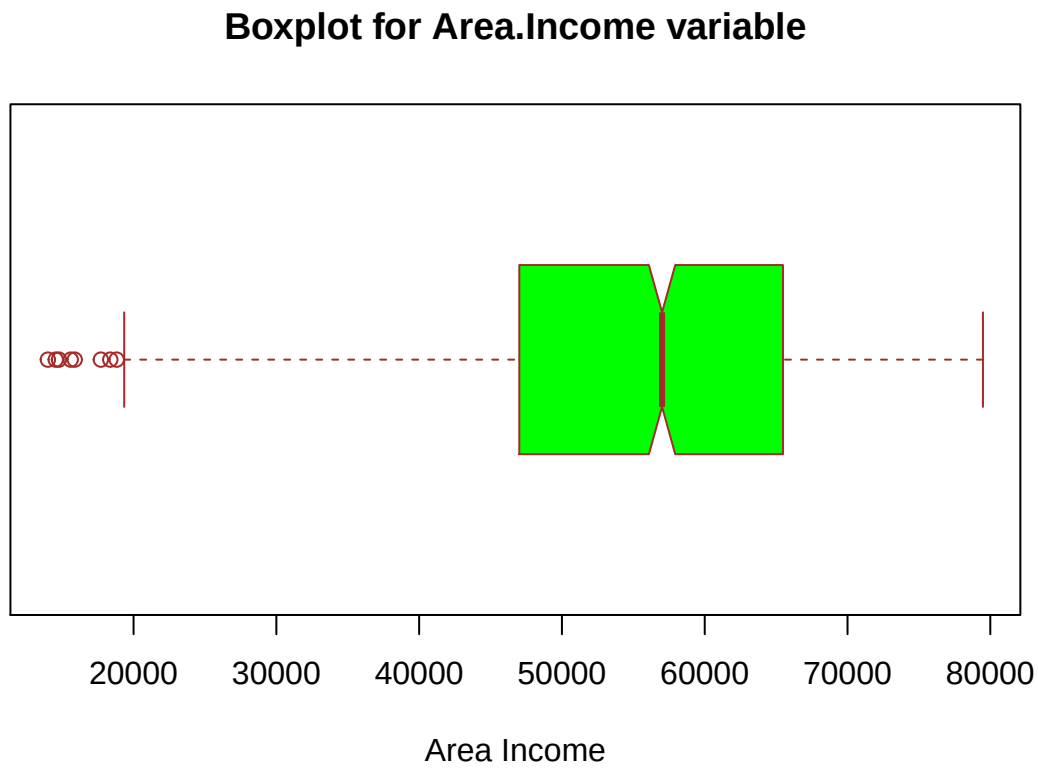
```
#We have 1000 rows and 10 columns
```

```
summary(df)
```

```
##  Daily.Time.Spent.on.Site      Age          Area.Income
##  Min.   :32.60            Min.   :19.00   Min.   :13996
##  1st Qu.:51.36            1st Qu.:29.00   1st Qu.:47032
##  Median :68.22            Median :35.00   Median :57012
##  Mean   :65.00            Mean   :36.01   Mean   :55000
##  3rd Qu.:78.55            3rd Qu.:42.00   3rd Qu.:65471
##  Max.   :91.43            Max.   :61.00   Max.   :79485
##
##  Daily.Internet.Usage                                 Ad.Topic.Line
##  Min.   :104.8        Adaptive 24hour Graphic Interface      :  1
##  1st Qu.:138.8        Adaptive asynchronous attitude         :  1
##  Median :183.1        Adaptive context-sensitive application :  1
##  Mean   :180.0        Adaptive contextually-based methodology:  1
##  3rd Qu.:218.8        Adaptive demand-driven knowledgebase   :  1
##  Max.   :270.0        Adaptive uniform capability            :  1
##                       (Other)                                :994
##               City          Male                Country
##  Lisamouth      :  3   Min.   :0.000   Czech Republic:  9
##  Williamsport   :  3   1st Qu.:0.000   France        :  9
##  Benjaminchester:  2   Median :0.000   Afghanistan   :  8
##  East John      :  2   Mean   :0.481   Australia     :  8
##  East Timothy   :  2   3rd Qu.:1.000   Cyprus        :  8
##  Johnstad       :  2   Max.   :1.000   Greece        :  8
##  (Other)        :986                   (Other)       :950
##                Timestamp   Clicked.on.Ad
##  2016-01-01 02:52:10:  1   Min.   :0.0
##  2016-01-01 03:35:35:  1   1st Qu.:0.0
##  2016-01-01 05:31:22:  1   Median :0.5
##  2016-01-01 08:27:06:  1   Mean   :0.5
##  2016-01-01 15:14:24:  1   3rd Qu.:1.0
##  2016-01-01 20:17:49:  1   Max.   :1.0
##  (Other)            :994
```

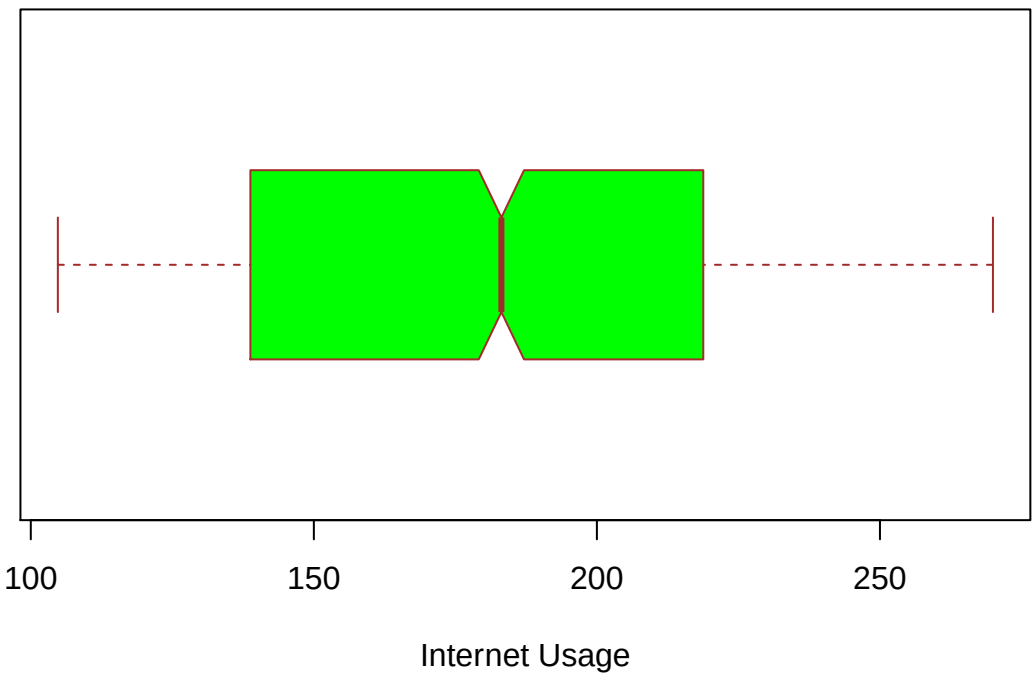
```
##DATA CLEANING
#1)Validity
#check outliers
#boxplot(df)
bxplt_Area.Income = boxplot(df$Area.Income,
```

```
                            main = "Boxplot for Area.Income variable",
                            xlab = "Area Income",
                            col = "green",
                            border = "brown",
                            horizontal = TRUE,
                            notch = TRUE
)
```



#Majority of the people earn more than 20000.However,we have those that lie outside the 20000

```
#in our dataset,we have high income earners
bxplt_Internet.Usage = boxplot(df$Daily.Internet.Usage,
                            main = "Boxplot for Daily.Internet.Usage",
                            xlab = "Internet Usage",
                            col = "green",
                            border = "brown",
                            horizontal = TRUE,
                            notch = TRUE
)
```




```
#There exists no outliers for daily internet usage
```

```
duplicated_rows <- df[duplicated(df),]
duplicated_rows
```

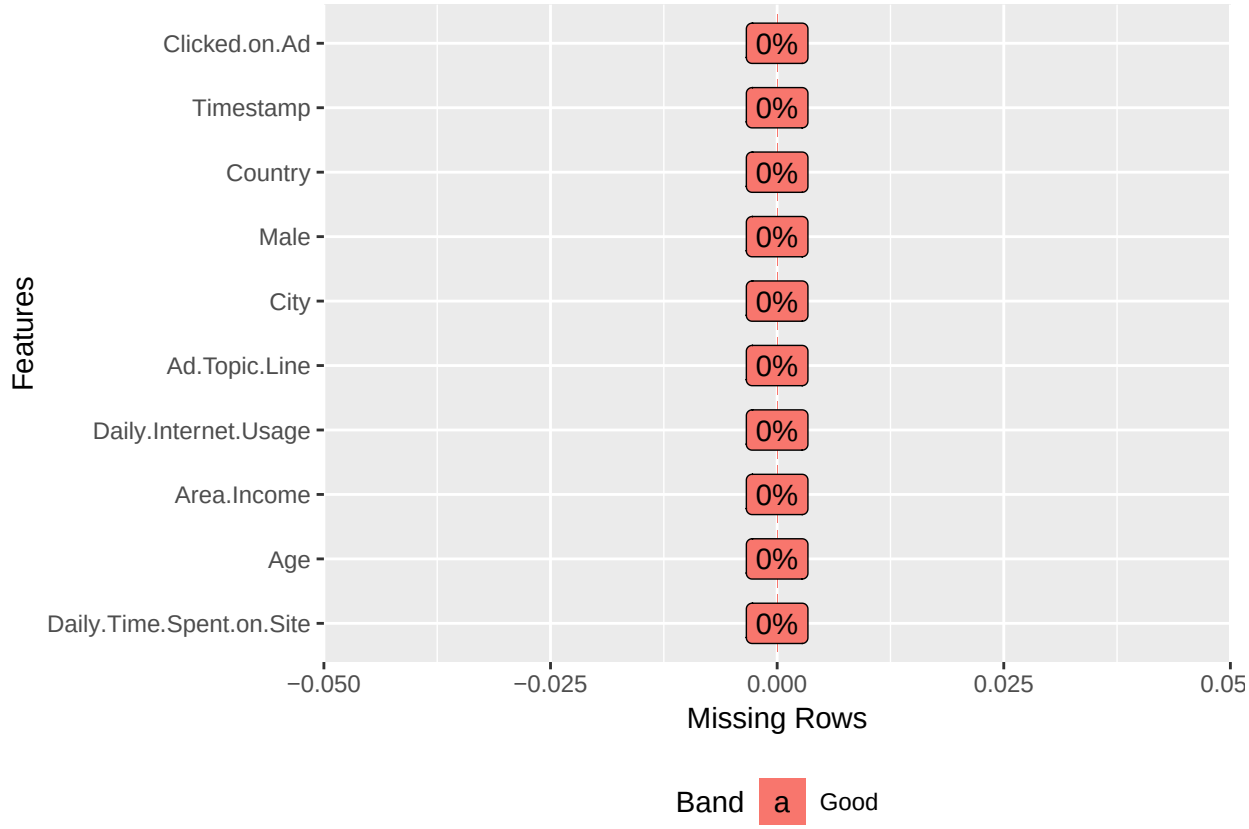
```
##  [1] Daily.Time.Spent.on.Site Age
##  [3] Area.Income              Daily.Internet.Usage
##  [5] Ad.Topic.Line            City
##  [7] Male                     Country
##  [9] Timestamp                Clicked.on.Ad
## <0 rows> (or 0-length row.names)
```

```
#there are no duplicated rows from the dataset
```

```
##UNIVARIATE ANALYSIS
library(DataExplorer)
```

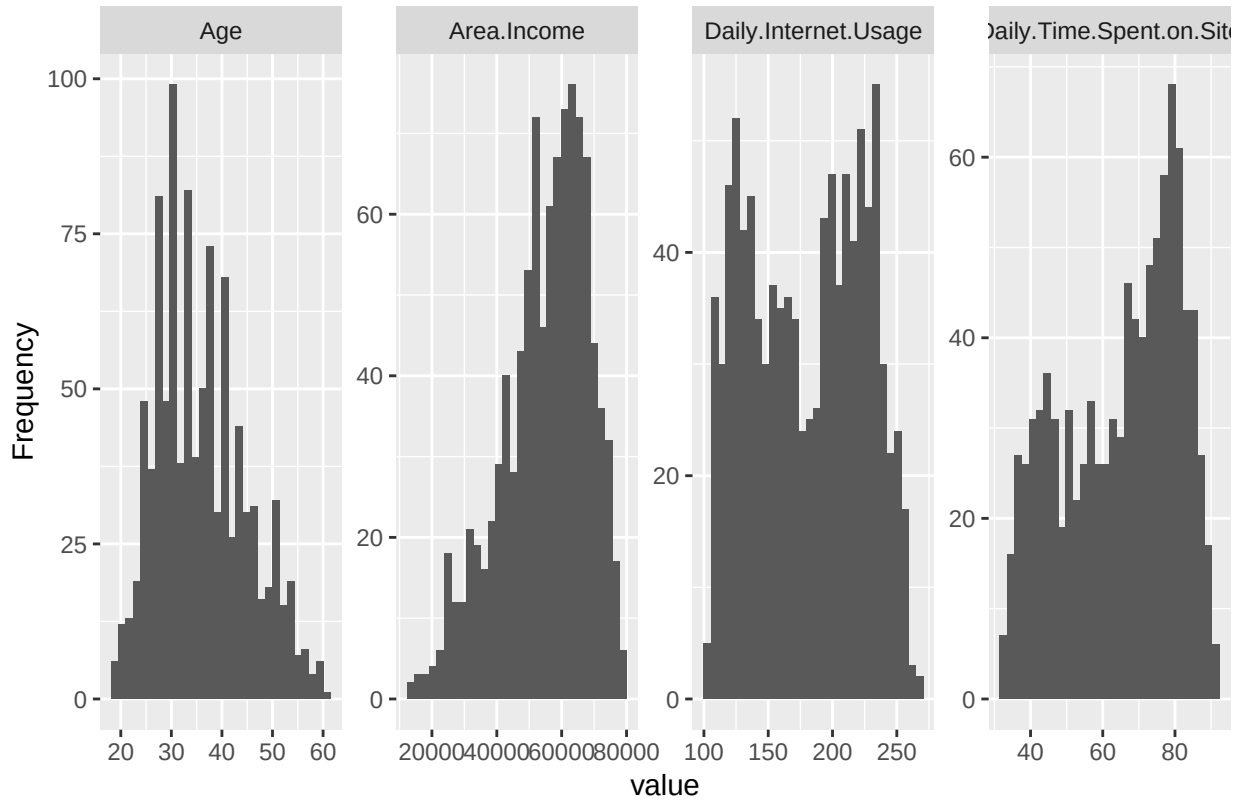
```
plot_str(df)
#Variables in the dataset and their data types
```

```
plot_missing(df)
```




```
#We do not have any missing values in the dataset
```

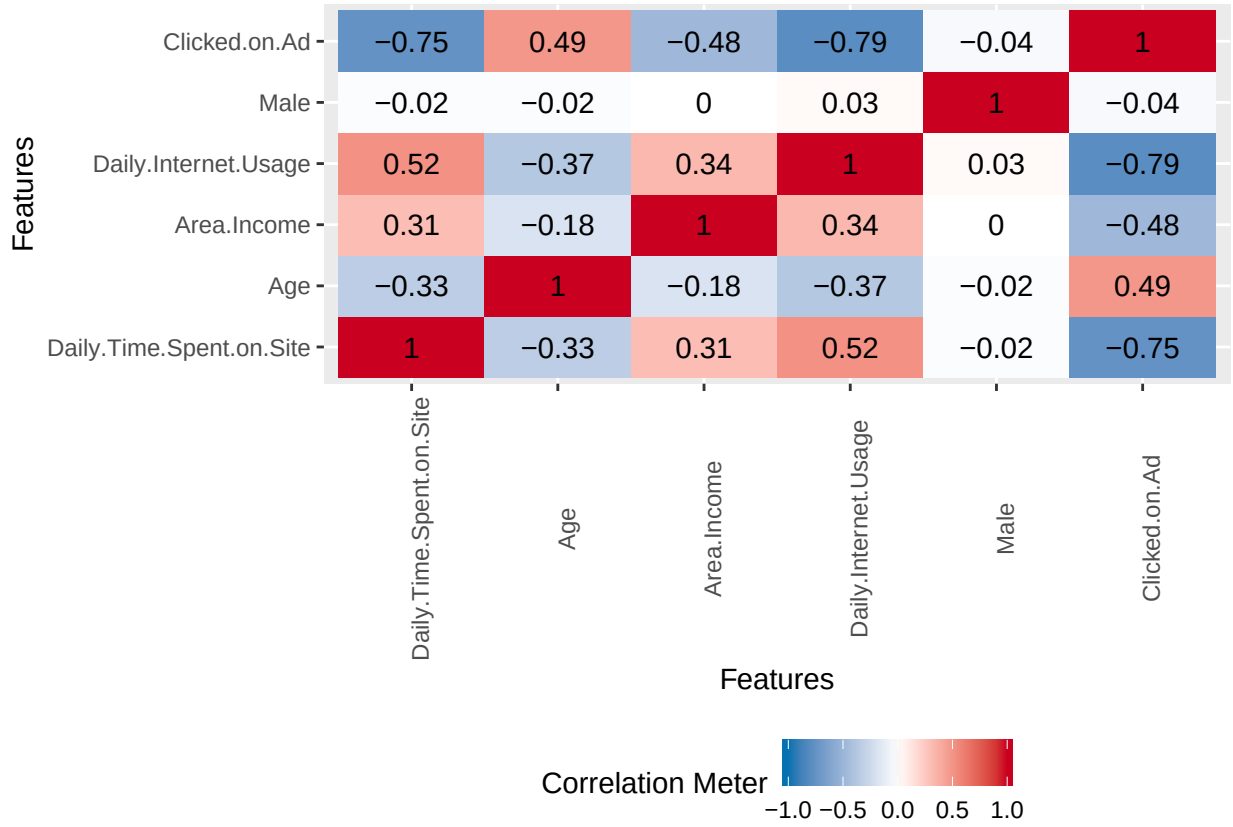
```
#Create graphs to see how the datset looks like
plot_histogram(df)
```




```
#majority of the people in the dataset are 30 years
```

Majority of the people are using the internet It is likely that they shall view the ad if it keeps popping Their income is also on the higher end (Our graph is skewed to the left hence symbolizing this) We have people who spend to much time on the ad.The more time you spend on the ad there is a high likelihood you will take the course

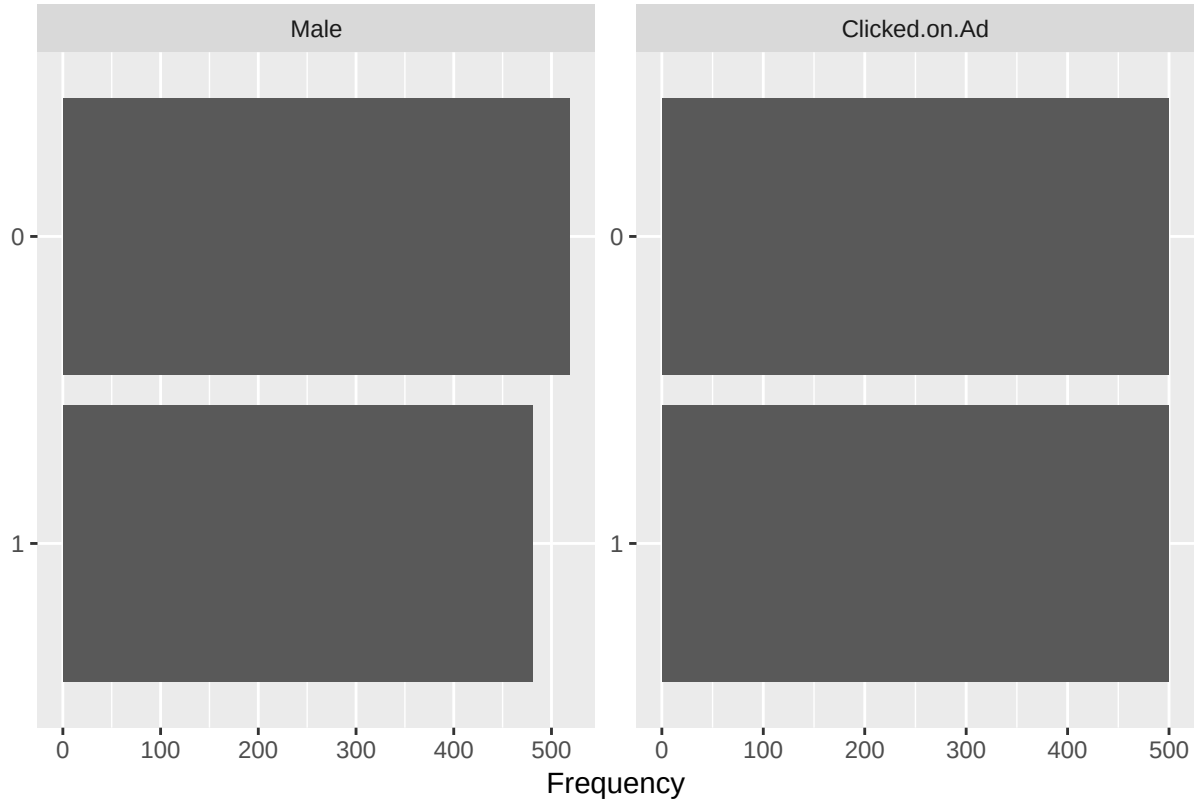
```
#Correlation
plot_correlation(df,type = 'continuous')
```



There exists a higher correlation between daily time spent onn the site and th data internet usage.This implies that,when the internet usage increases,you are more likely yo pend more time on the site. still,there exists weak correlation between some variables. This means,when on variable increases the othe variable will decrease

```
plot_bar(df)
```

```
## 4 columns ignored with more than 50 categories.
## Ad.Topic.Line: 1000 categories
## City: 969 categories
## Country: 237 categories
## Timestamp: 1000 categories
```



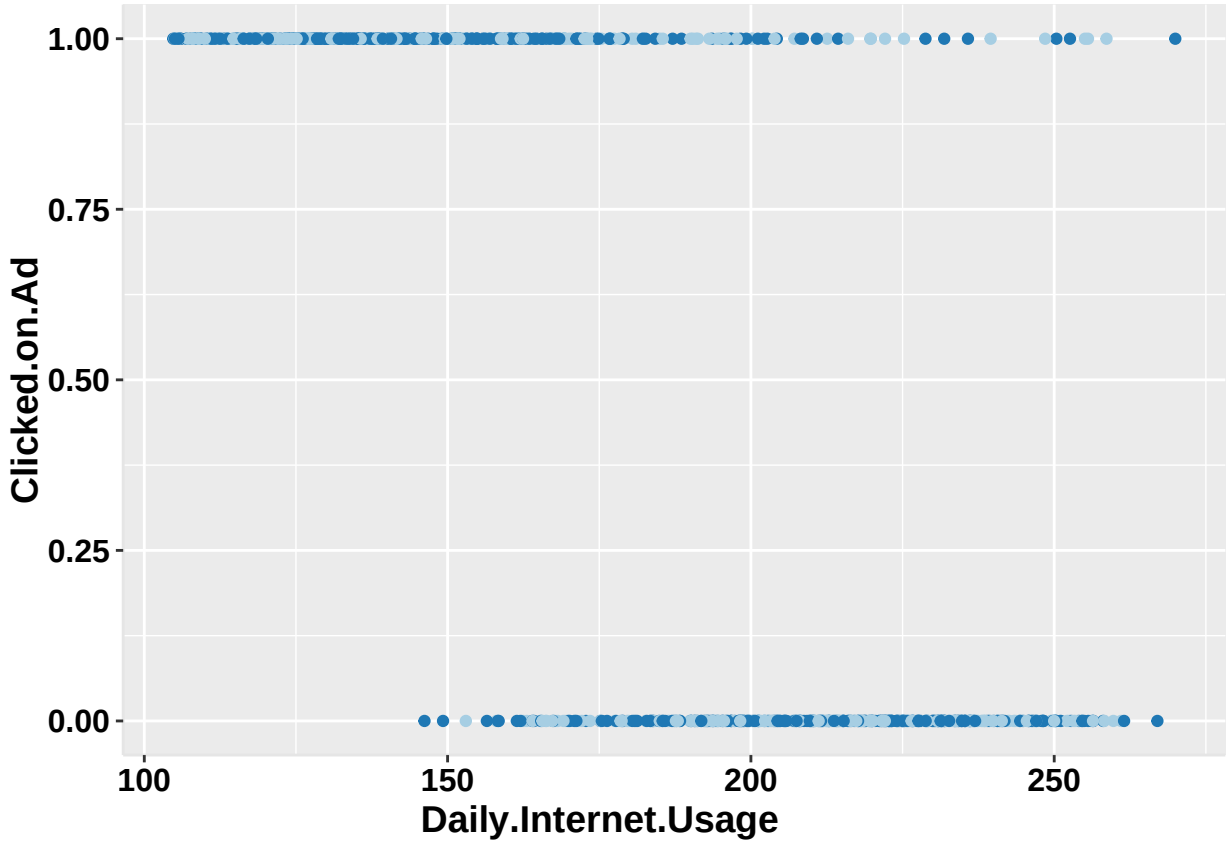
There are more females in the data collected compared to the males There was an equal number of people who clicked on the ad and those who did not click on the ad

```
#call the library
library(devtools)
```

```
library(easyGgplot2)
```

```
## Loading required package: ggplot2
```

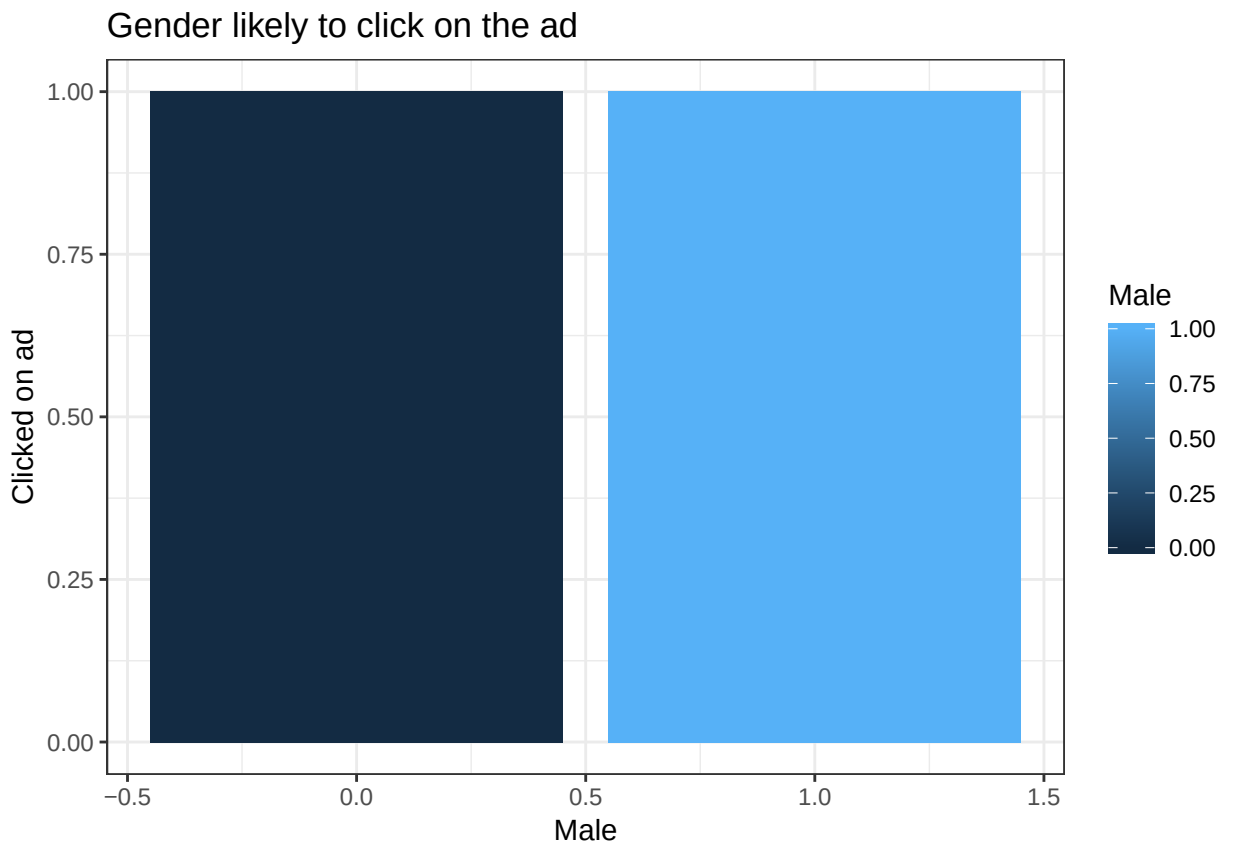
```
#Comparing if they clicked on ad based on the internet usage
ggplot2.scatterplot(data =df,xName = 'Daily.Internet.Usage', yName = 'Clicked.on.Ad',groupName = 'Male'
```



I could say majority of the people who spend less time on the internet are likely to click on the ad

```
df2<-df
```

```
p <-ggplot(df2, aes(Male, Clicked.on.Ad))
p +geom_bar(stat = "identity", aes(fill = Male), position = "dodge") +
  xlab("Male") + ylab("Clicked on ad") +
  ggtitle("Gender likely to click on the ad") +
  theme_bw()
```




```
#Either male and female are likely to click on the ad
```

```
df2[,'Clicked.on.Ad']<-factor(df2[,'Clicked.on.Ad'])
```

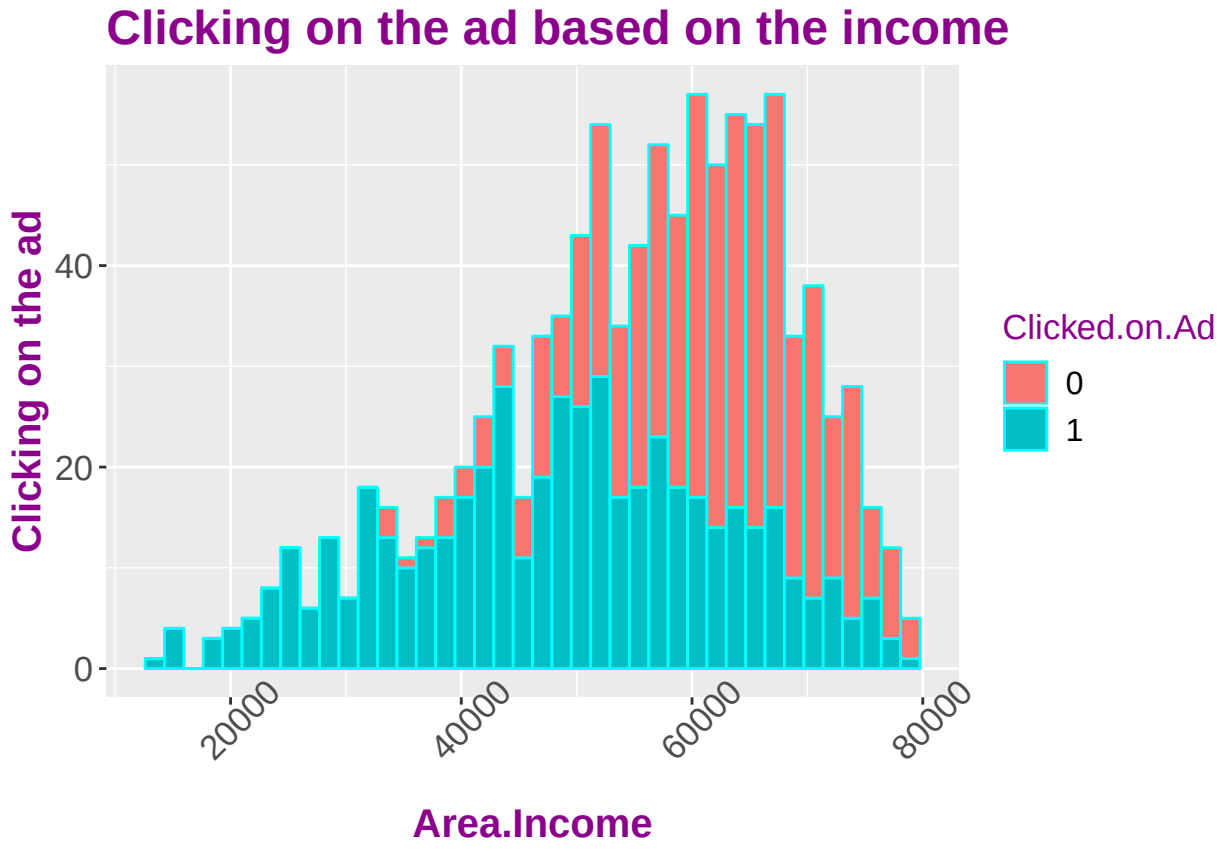
```
head(df2)
```

```
##   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
```

clicked on ad has been converted to a factor

```
options(repr.plot.width = 13, repr.plot.height = 7)
ggplot(data = df2, aes(x = Area.Income, fill = Clicked.on.Ad))+
    geom_histogram(bins = 40, color = 'cyan') +
    labs(title = 'Clicking on the ad based on the income', x = 'Area.Income', y = 'Clicking on the ad',
        scale_color_brewer(palette = 'Paired') +
        theme(plot.title = element_text(size = 18, face = 'bold', color = 'darkmagenta'),
             axis.title.x = element_text(size = 15, face = 'bold', color = 'darkmagenta'),
             axis.title.y = element_text(size = 15, face = 'bold', color = 'darkmagenta'),
             axis.text.x = element_text(size = 13, angle = 45),
             axis.text.y = element_text(size = 13),
             legend.title = element_text(size = 13, color = 'darkmagenta'),
             legend.text = element_text(size = 12))
```

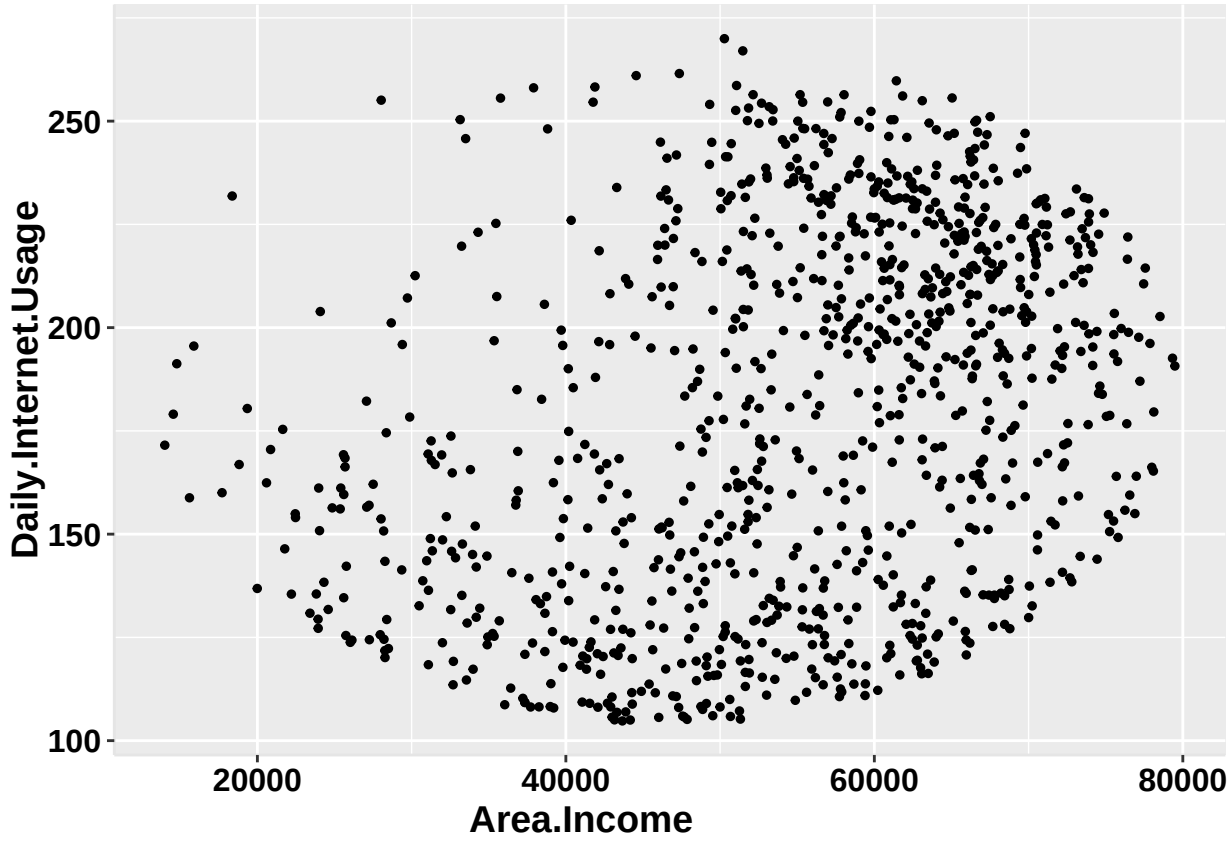


Most of the low earners will click on the ad It could be like this since they are probably hoping to get more knowledge to improve their lifestyle It can also be because they have more free time

```
df2$Daily.Internet.Usage <- factor(df2$Daily.Internet.Usage)
```

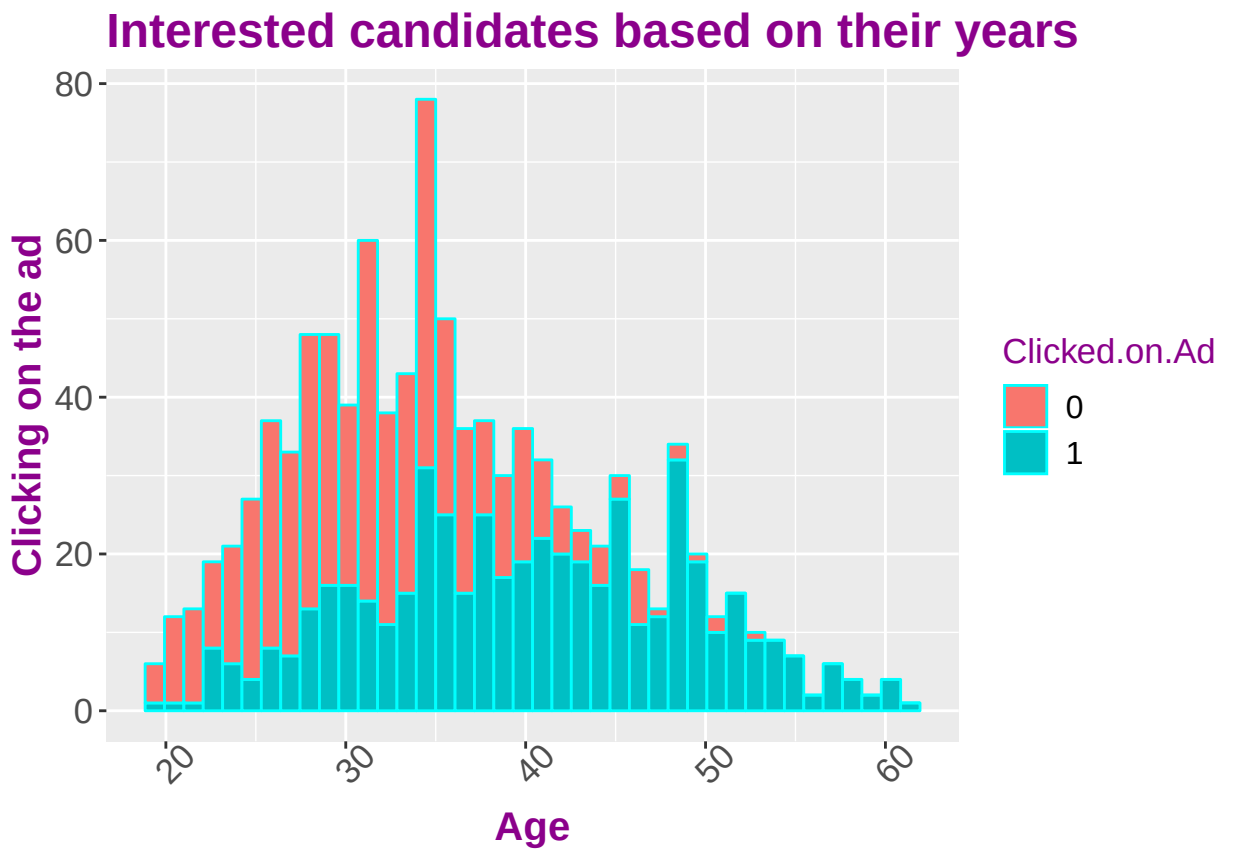
```
library(ggplot2)
```

```
ggplot2.scatterplot(data=df, xName='Area.Income',yName='Daily.Internet.Usage'
                    )
```



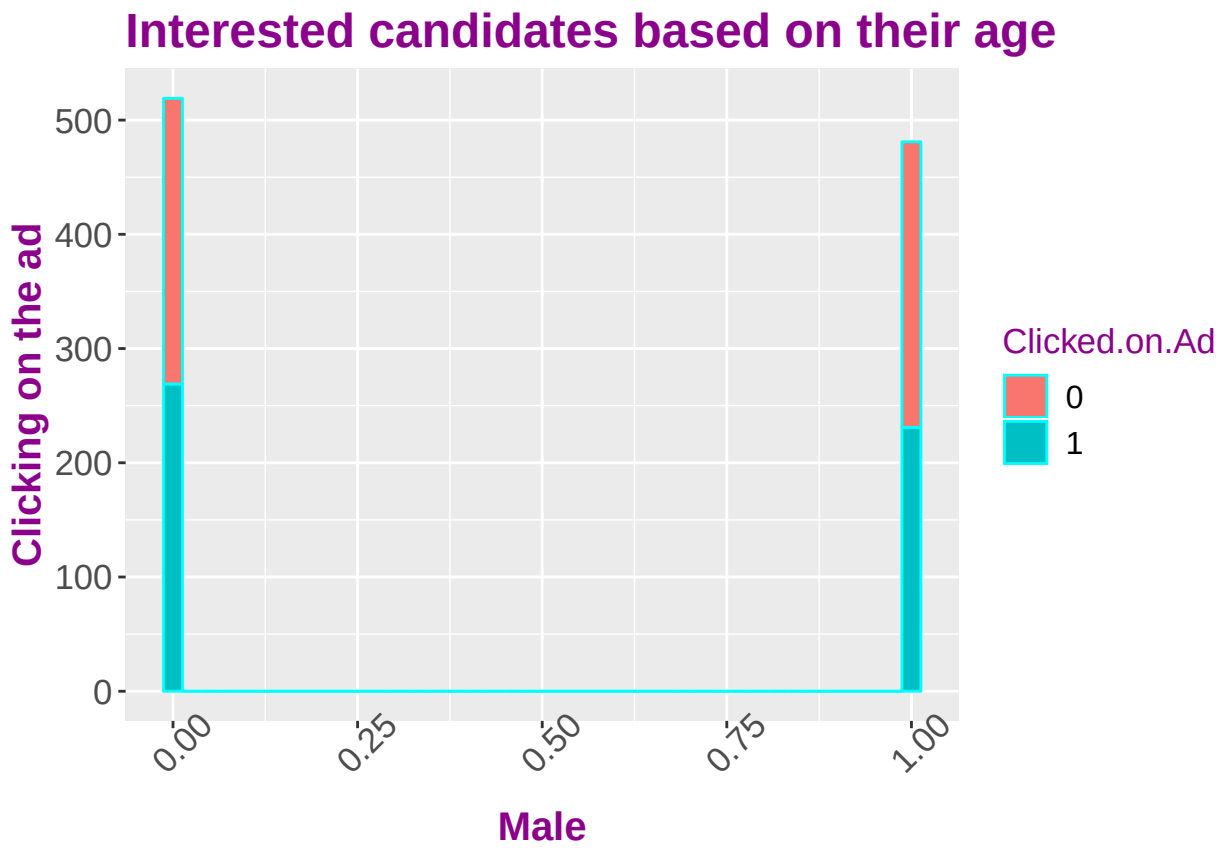
This proves that low income earners spend less time on the internet while the high income earners spend more time on the internet

```
#Create a stacked graph that will show ad clicked by people of different ages
options(repr.plot.width = 13, repr.plot.height = 7)
ggplot(data = df2, aes(x = Age, fill = Clicked.on.Ad))+
    geom_histogram(bins = 40, color = 'cyan') +
    labs(title = 'Interested candidates based on their years', x = 'Age', y = 'Clicking on the ad', fil
        scale_color_brewer(palette = 'Paired') +
        theme(plot.title = element_text(size = 18, face = 'bold', color = 'darkmagenta'),
             axis.title.x = element_text(size = 15, face = 'bold', color = 'darkmagenta'),
             axis.title.y = element_text(size = 15, face = 'bold', color = 'darkmagenta'),
             axis.text.x = element_text(size = 13, angle = 45),
             axis.text.y = element_text(size = 13),
             legend.title = element_text(size = 13, color = 'darkmagenta'),
             legend.text = element_text(size = 12))
```



#Older people are more likely to click of the ad compared to younger people #This could be due to them having financiall access to pay the course for themselves

```
#Create a stacked graph that will show ad clicked by people of different ages
options(repr.plot.width = 13, repr.plot.height = 7)
ggplot(data = df2, aes(x = Male, fill = Clicked.on.Ad))+
    geom_histogram(bins = 40, color = 'cyan') +
    labs(title = 'Interested candidates based on their age', x = 'Male', y = 'Clicking on the ad', fill
        scale_color_brewer(palette = 'Paired') +
        theme(plot.title = element_text(size = 18, face = 'bold', color = 'darkmagenta'),
             axis.title.x = element_text(size = 15, face = 'bold', color = 'darkmagenta'),
             axis.title.y = element_text(size = 15, face = 'bold', color = 'darkmagenta'),
             axis.text.x = element_text(size = 13, angle = 45),
             axis.text.y = element_text(size = 13),
             legend.title = element_text(size = 13, color = 'darkmagenta'),
             legend.text = element_text(size = 12))
```



There is an equal chance of clicking on the ad for both males and females

#Conclusion 1) Majority of the people on the internet are high earners. However,they are not interested in the cryptography course. 2) Most of the people interested are older 3) Spending more time on the internet does not necessarily mean they will click on the ad

#Recommendation 1) Since most of the interested are low earners,it would be wise to have the cost of the course affordable to most people This will lead to high number joining the course hence the profits will increase. 2) Educate the younger people on the importance of the course.They probably are not clickingon it since they are either ignorant are not not interested in general