Moringa_Dsc14_Core_Week12_IP_Rprogramming_Jonah_Okiru_05_2022

Jonah okiru

2022-05-27

##1. Define question.

#a).Specifying the question.

To help an online cryptography entrepreneur to identify the individuals who are most likely to click on her advert.

#b). The metric of success.

To identify individual who are most likely to click on an advert with an accuracy score of 90%

#c). The context

With the ever growing in the technology globally and internally, it has make it possible for entrepreneurs to market their products on global scale while at the comfort of their homes. These has been successful, thanks to platforms such as blogs.These platforms attracts viewers and readers of the blogs content from all over the world and as they do so , adverts of various types pops up after certain time intervals, Some of the visitors of these platforms click on these adverts that pops up on the platform as they consume the content of the platform. The owners of these platforms always would like to know the type of individuals who are visitors to the platforms and are likely to click on the pop up adverts so that they could promote the products which target those individuals at their platforms.

#d). recording Exprimental design

Hypothesis: To identify individual who visit the blog platform and the individual is most likely to click on the pop up adverts.

X-Axis: Age, gender, income area, city and country.

Y-Axis: Click on advert.

Experimental set up: I identify the individual who visits the blog platform as either most likely to click on the advert or not.

Design of the experiment: Analyze the 1000 records from the previous visitors of the blog platform.

Sample size: 1000 records from the previous visitors of the blog.

```
#Load the data
library("data.table")
advertising <- fread("http://bit.ly/IPAdvertisingData")
#Print the first six records of the data
head(advertising)
```

```
##    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
```

```r
#Drop unnecessary columns.
#select the variables with names of columns to be retained
age <- advertising$Age
gender <- advertising$Male
Click_on_advert <- advertising$`Clicked on Ad`
area_income <- advertising$`Area Income`
city <- advertising$City
country <- advertising$Country

#Create the data table from the selected variables and assigned it to df
df <- data.table(age, gender, area_income, city, country, Click_on_advert)

#Print the first six records of the data
head(df)
```

```
##    age gender area_income           city     country Click_on_advert
## 1:  35      0    61833.90    Wrightburgh     Tunisia               0
## 2:  31      1    68441.85      West Jodi       Nauru               0
## 3:  26      0    59785.94       Davidton San Marino               0
## 4:  29      1    54806.18 West Terrifurt       Italy               0
## 5:  35      0    73889.99   South Manuel     Iceland               0
## 6:  23      1    59761.56      Jamieberg      Norway               0
```

The other columns were dropped in the cell above since they're of no importance in the further analysis of the data hence had to be dropped.

```r
#Check for the missing value in the
#Find the total missing value in each column of the df dataset.
colSums(is.na(df))
```

```
##             age          gender     area_income            city         country
##               0               0               0               0               0
## Click_on_advert
##               0
```

All the column of the dataset has no missing values.

```r
#Check for the outliers within the dataset numeric columns using boxplot.
#Check for the existence of outliers in the age column
age <- df$age
#The box plot
boxplot(age)
```
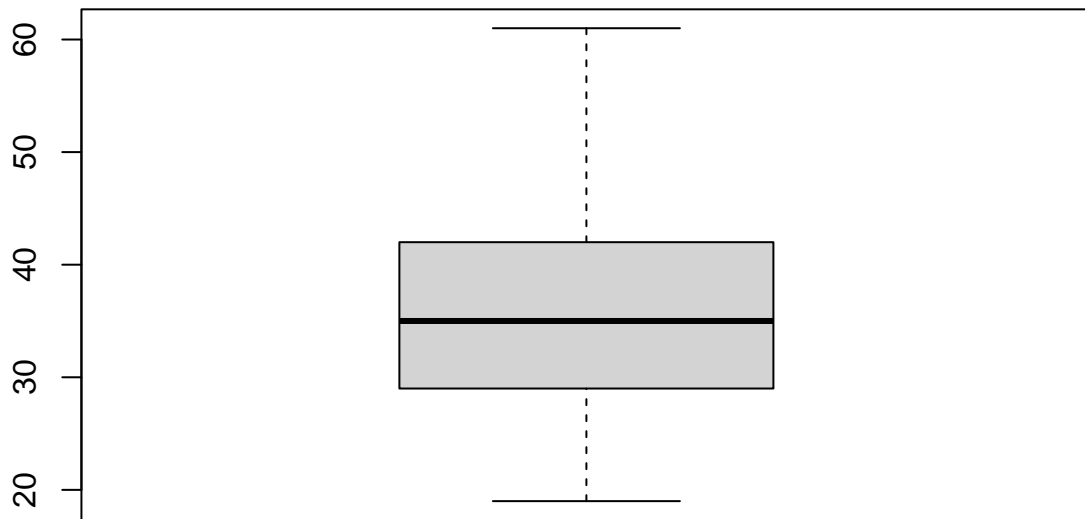

From the above age column boxplot, we note that the age column data has no outliers.

```r
#Check for the existence of outliers in the gender column
Gender <- df$gender
#Boxplot of gender
boxplot(Gender)
```
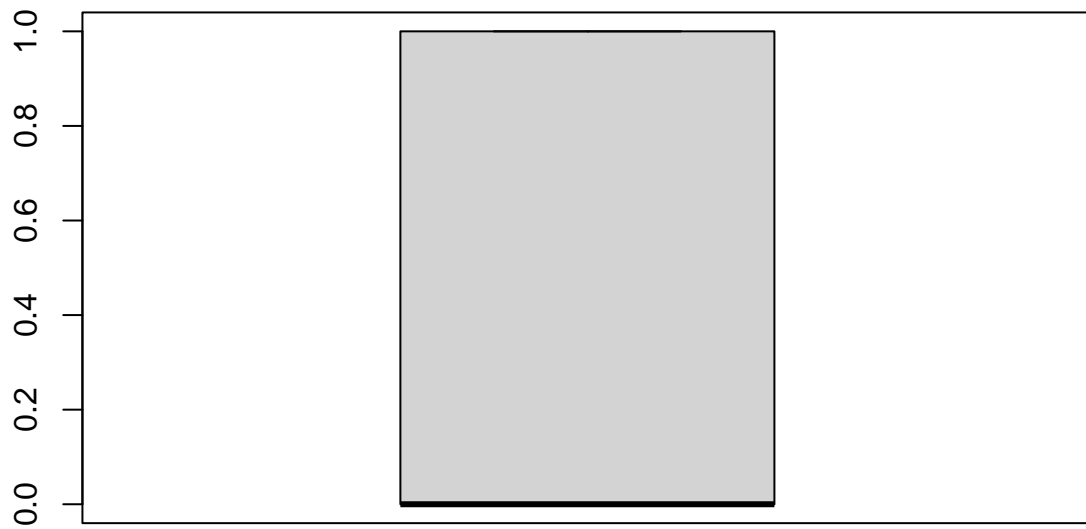
From the boxplot above , we conclude that there is no outliers in the gender column. This is due to values in the gender columns being categorical so they are either class '0' or class one '1'

```r
#Check for the existence of outliers in the income column
area_income <- df$area_income
#Box plot
boxplot(area_income)
```
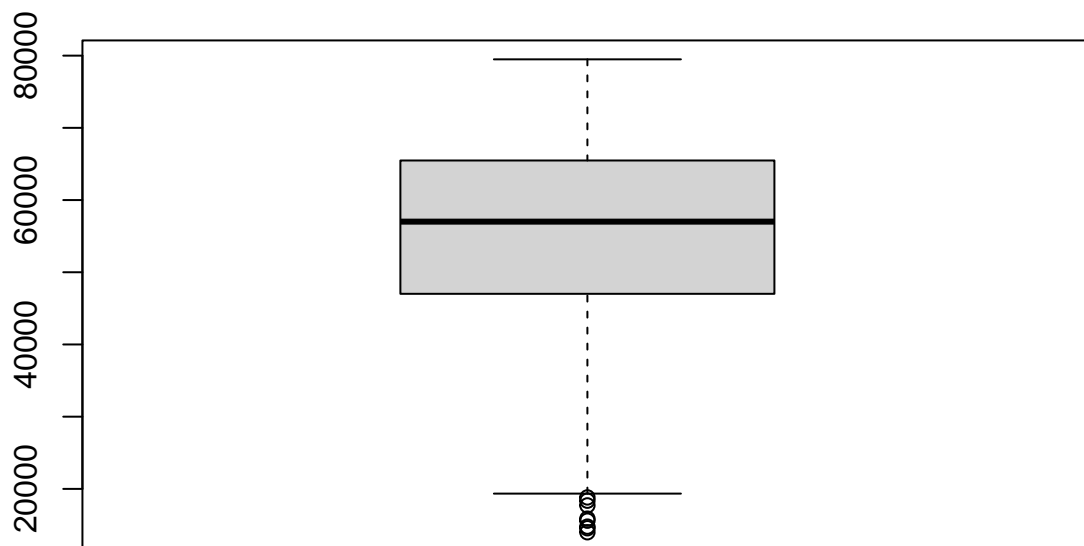
The existence of outliers in the income column is due disparities of income level among countries and cities.

```r
#Check for duplicates in the dataset
duplicate_rows <- df[duplicated(df),]
#print the duplicate rows
duplicate_rows
```

```
## Empty data.table (0 rows and 6 cols): age,gender,area_income,city,country,Click_on_advert
```

There were no duplicate records in the dataset.

```r
#Univariate analysis
#Check for the descriptive statistics of the data
library(descr)
descr(df)
```

```
##       age              gender          area_income          city
##  Min.   :19.00    Min.    :0.000    Min.    :13996    Length:1000
##  1st Qu.:29.00    1st Qu.:0.000    1st Qu.:47032    Class :character
##  Median :35.00    Median :0.000    Median :57012    Mode  :character
##  Mean   :36.01    Mean    :0.481    Mean    :55000
##  3rd Qu.:42.00    3rd Qu.:1.000    3rd Qu.:65471
##  Max.   :61.00    Max.    :1.000    Max.    :79485
##    country          Click_on_advert
##  Length:1000       Min.    :0.0
##  Class :character  1st Qu.:0.0
```

```
##  Mode  :character    Median :0.5
##                       Mean   :0.5
##                       3rd Qu.:1.0
##                       Max.   :1.0
```

From the above descriptive statistics of the columns, we observe that the minmum age of was 19 years, maximum age was 61 years, the mean age is 36 years and the median age is 35 years. Also we note that the minimum area income is 13996, the mean area income is 55000, the median is 57012 and the maximum is 79485

```
#Barplots of categorical class columns
#Bar plot of gender
#Frequency table of gender column
gender_freq_table <- table(gender)
#Barplot of gender
barplot(gender_freq_table)
```
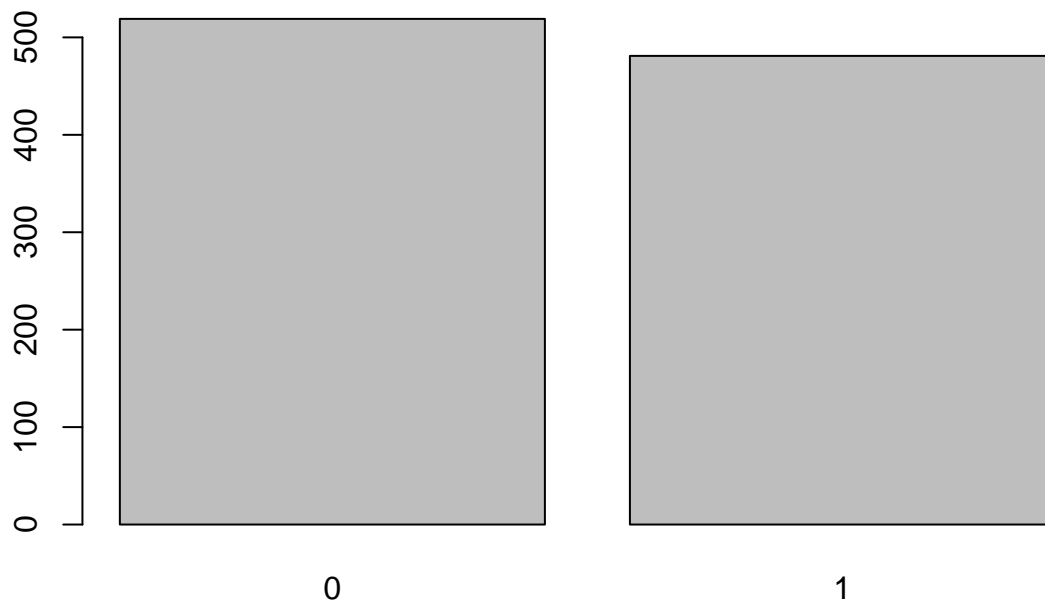

From the bar chart we note the number of female gender who click on advert is slightly higher compared to the male gender

```
#Barplot of city
#Frequency table of the city
city_freq_table <- table(city)
#Barplot of the city
barplot(city_freq_table)
```
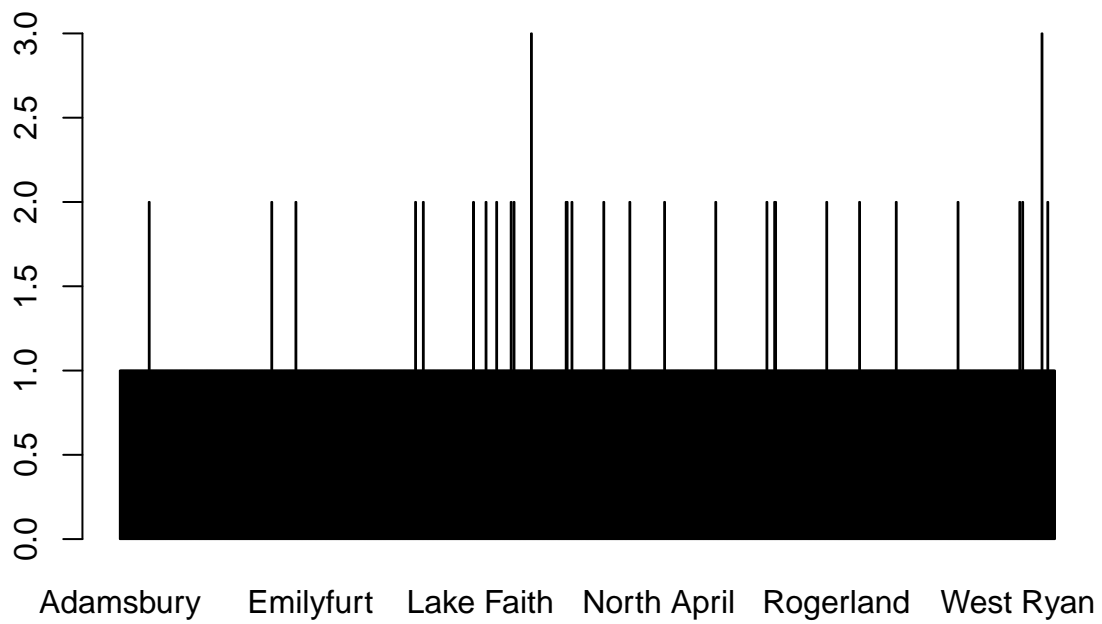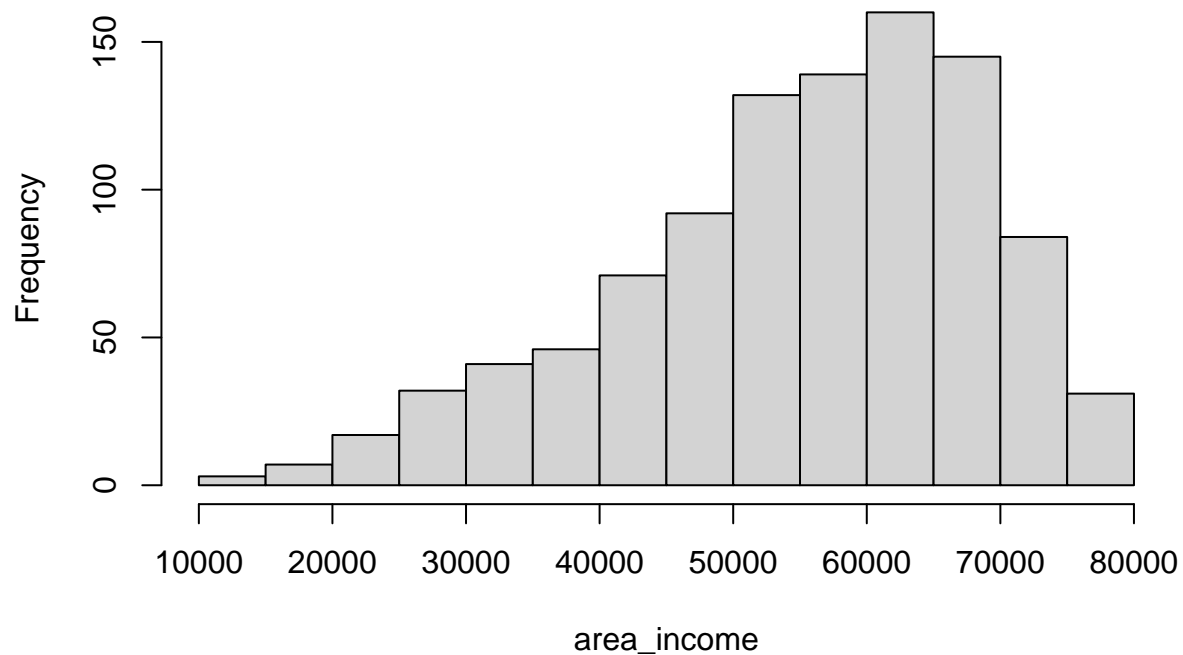
```
#Bar plot of the country
#Country frequency table
country_frequencytable <- table(country)
#Barplot of country column
barplot(country_frequencytable)
```
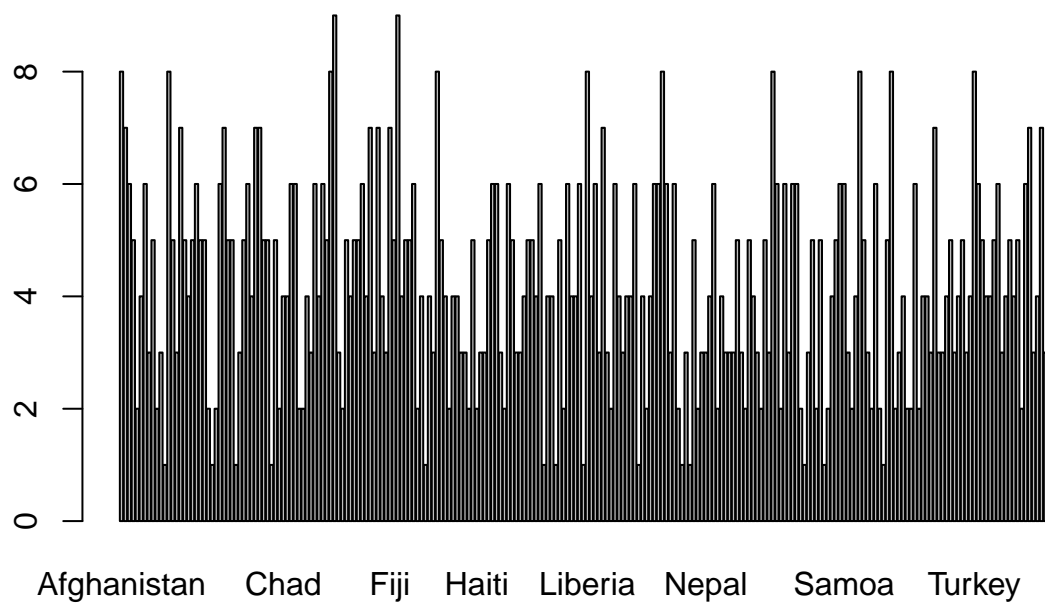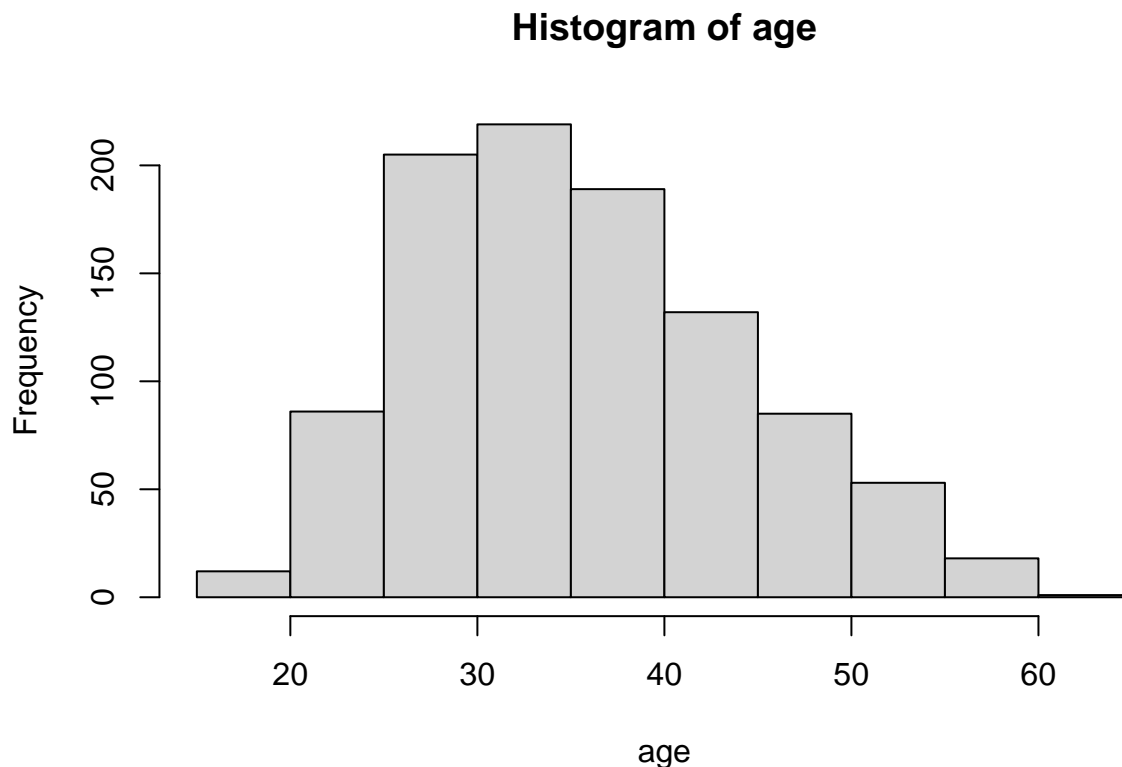
```
#Plot the Histograms
#The histogram of area income
area_income <- df$area_income
hist(area_income)
```

Histogram of area_income

From the histogram above, it's bit skewed the left these implies that majority of the people have area income less than the mean area income.

```
#Histogram of age column
age <- df$age
hist(age)
```

Histogram of age

From the histogram it's noted that age histogram is skewed to the right these implies that the age of majority of the participants in the datasets was above the mean age.

```
#Bivariate and multivariate analysis
#Covariance between age and click on advert
#Age variable
age <- df$age
#click on the advert variable
Click_on_advert <- df$Click_on_advert
#covariance between age and click on advert
cov(age, Click_on_advert)
```

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

The above covariance of 2.164665 implies that the relationship between the age and click on advert is a positive relationship. These means that a when the age increases the the click on advert also increases and vice versa.

```
#Covariance between gender and click on advert
gender <- df$gender
#Covariance btwn gender and click on advert
cov(gender, Click_on_advert)
```

```
## [1] -0.00950951
```

The covariance of -0.00950951 above indicates there is very slight negative relationship between the gender and the click on advert. These implies there is no a noticeable effect when the genders changes on the click on advert.

```
#Covariance between area income and click on advert
area_income <- df$area_income
cov(area_income, Click_on_advert)
```

```
## [1] -3195.989
```

The covariance above of -3195.989 implies that the relationship between area income and click on advert is a negative relationship. These implies that a slight decrease in the income will lead to a huge increase on the the click on advert.

```
#Correlation between gender and click on advert
gender <- df$gender
#Covariance btw gender and click on advert
cor(gender, Click_on_advert)
```

```
## [1] -0.03802747
```

The correlation above of -0.03802747 , indicates that the correlation between the gender and click on advert is very weak negative correlation.

```
#Correlation between gender and click on advert
age <- df$age
#Covariance btwn gender and click on advert
cor(age, Click_on_advert)
```

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

The correlation above 0.4925313, implies that the correlation between age and click on advert is a moderate positive correlation. These implies an increase in age leads to an increase on the click on advert at the same proportion.

```
#Correlation between area income and click on advert
area_income <- df$area_income
cor(area_income, Click_on_advert)
```

```
## [1] -0.4762546
```

The correlation above of -0.4762546 is a moderate negative correlation. These implies that a decrease in area income leads to proportionate increase in the click on the advert.

```
#Scatter plot between gender and click on advert
plot(Click_on_advert, gender, xlab=("gender"), ylab =("click on advert"), main = "scatterplot of click
```
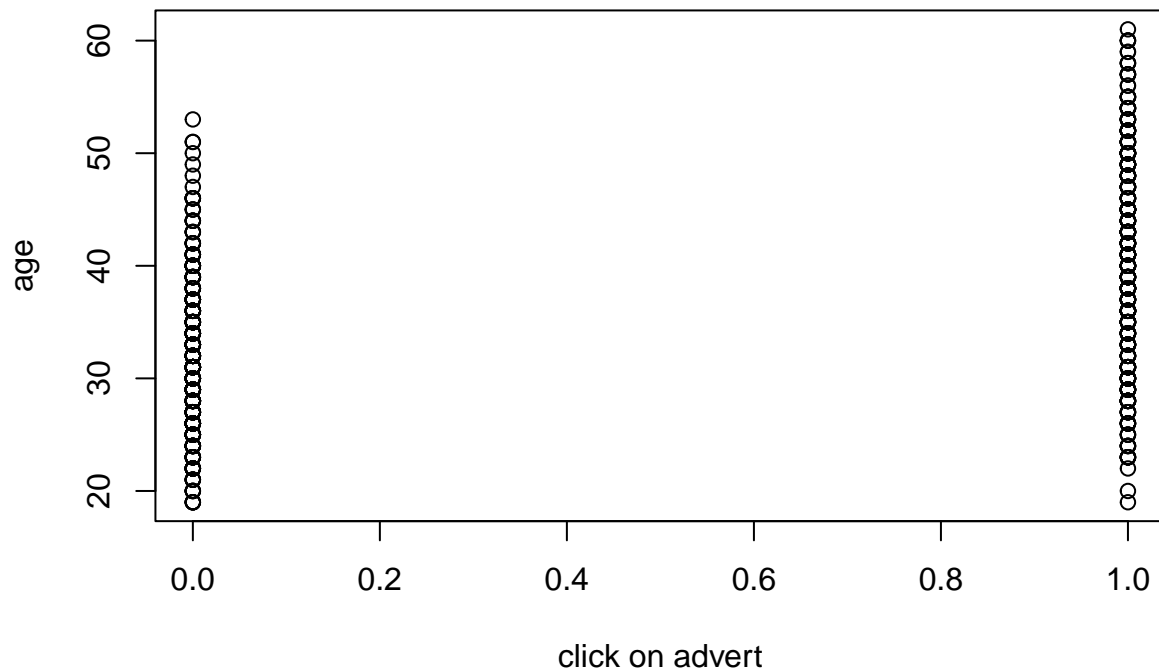
scatterplot of click on advert vs gender
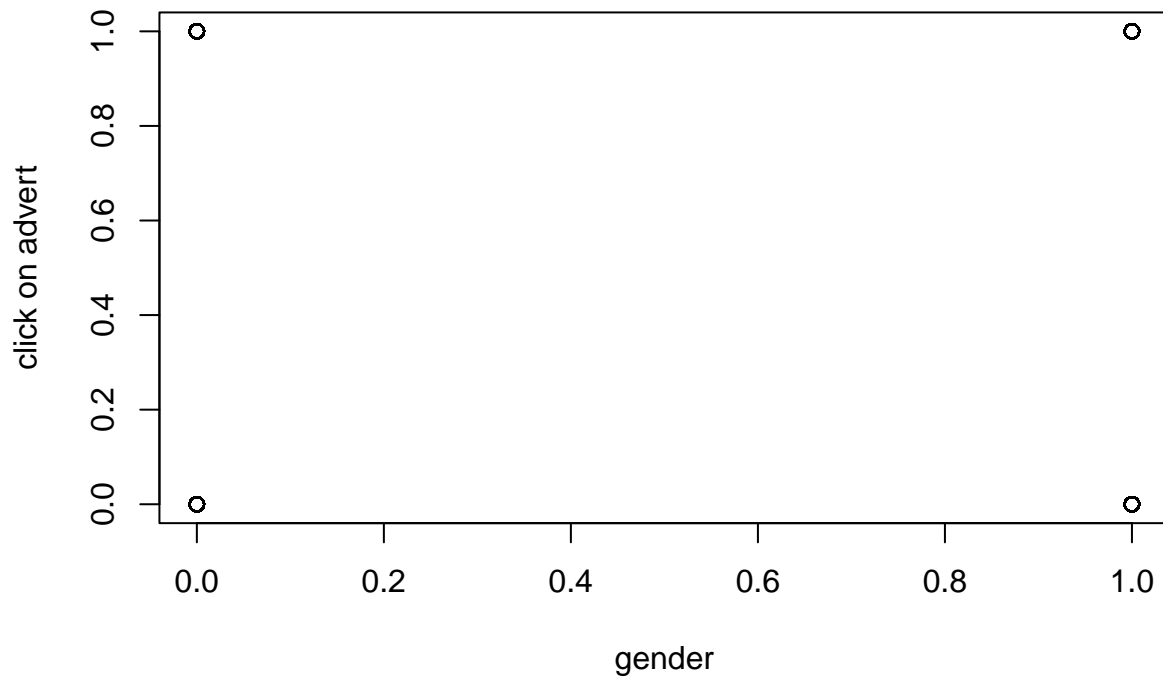
The scatter plot above between the gender and the click on advert implies that there's no correlation between the gender and click on the advert. A change in gender will not lead to change on the click on advert and vice versa.

```
#Scatter plot btwn age and click on advert
plot(Click_on_advert, age, xlab = "click on advert", ylab = "age", main = "scatterplot of click on adve
```

scatterplot of click on advert vs age

The scatter plot above between age and click on advert shows there's a moderate positive relationship between gender and click on advert. These implies a positive change in age will also leads to a proportionate increase on the clicks on advert.

```
#Scatter plot between area income and click on advert
plot(Click_on_advert, area_income, xlab="click on advert", ylab = "area income", main= "scatterplot of
```

scatterplot of click on adver vs income
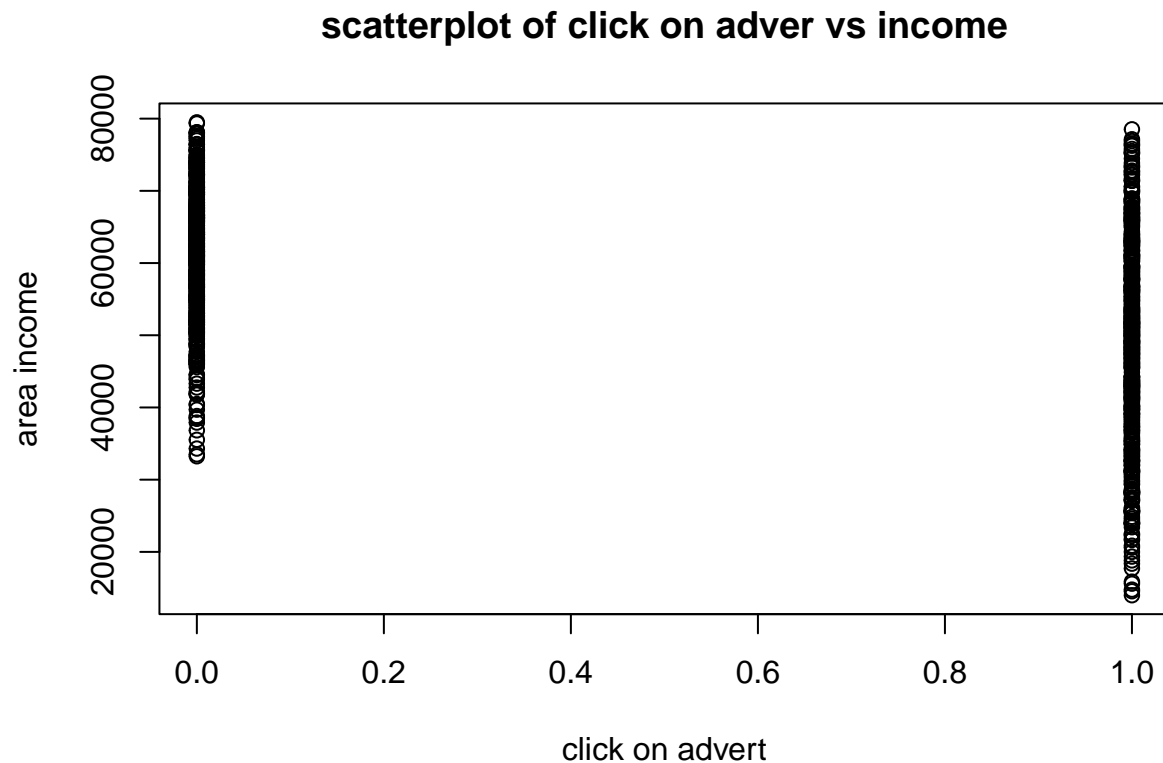
The scatter plot above between the area income and click on advert indicates the relationship between the area income and click on advert is a moderate negative relationship. These implies that positive change in area income will lead to a negative change in the click on the advert.

CONCLUSION

1. There's no any relationship between the gender of an individual who visits the blog and the chances of he or she clicking the adverts on the blog site

2.There's a moderate positive relationship between the age of individuals who visit the blog site and the chances of the visitor to click on the advert on the blog site

3. There's a moderate negative relationship between individual area income whom visits the blog site and the chances of the individual to vlick on the advert at the blog site.

Recommendation

The individual who is most likely to click on the advert on blog site, is person of gender either male or female whose area national income is less than 55000 and the age is greater than the 36 years.

Challenging the solution

The conclusion and recommendations have been made based on the univariate and bivariate analysis rather than the predictive model, these leaves some room of doubt on the efficient of the final result on which the conclusion was drawn from.

Follow up question

a). Do we have the right data?

Yes, the data was appropriate with no missing values.

b). Do we need another data?

No, the data was appropriate. The only thing needed is further analysis.

c).Do we have the right question?

Yes, the question is clear and straight forward.