# Week 12 Independent Project

## Defining The Question

### a) Specifying the Question.

Identify which individuals are most likely to click on her ads in Cryptography course website.

### b) Defining the Metrics for Success.

For this study, we will perform conclusive Exploratory Data Analysis to enable us identify individuals who are most likely to click on ads.

### c) Understanding the context.

A Kenyan entrepreneur has created an online cryptography course and would want to advertise it on her blog. She currently targets audiences originating from various countries. In the past, she ran ads to advertise a related course on the same blog and collected data in the process. Using the data previously collected, she is looking to do a study to identify which individuals are most likely to click on her ads.

### d) Experimental Design.

The project was undertaken using the following design Datasets(<http://bit.ly/IPAdvertisingData>)

* Load dataset
* Data Cleaning
* Performing Exploratory Analysis
* Conclusion

### e) Data Relevance

Data is provided was collected in the past but from the same blog hence it is very suitable for this study.

Definition of Variables Daily Time Spent on Site

Age

Area

Income

Daily Internet Usage

Ad Topic Line

City

Male

Country

Timestamp

Clicked on Ad

## Data Preparation

### Importing the Libraries

# load libraries  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.1.0 v dplyr 1.0.5  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

options(warn = -1)

### Loading the data

#load data  
df <- fread('http://bit.ly/IPAdvertisingData')

#priview the first 6 rows  
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

# Checking the class of our dataset  
class(df)

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

# Find the number of column and rows  
class(df); ncol(df); nrow(df)

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

## [1] 10

## [1] 1000

We have 10 rows and 1000 columns

#checking the first 10 rows in our dataset  
(print(head(df, n=10)))

## 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  
## 7: 88.91 33 53852.85 208.36  
## 8: 66.00 48 24593.33 131.76  
## 9: 74.53 30 68862.00 221.51  
## 10: 69.88 20 55642.32 183.82  
## 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  
## 7: Enhanced dedicated support Brandonstad 0 Myanmar  
## 8: Reactive local challenge Port Jefferybury 1 Australia  
## 9: Configurable coherent function West Colin 1 Grenada  
## 10: Mandatory homogeneous architecture Ramirezton 1 Ghana  
## 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  
## 7: 2016-01-28 20:59:32 0  
## 8: 2016-03-07 01:40:15 1  
## 9: 2016-04-18 09:33:42 0  
## 10: 2016-07-11 01:42:51 0

## 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  
## 7: 88.91 33 53852.85 208.36  
## 8: 66.00 48 24593.33 131.76  
## 9: 74.53 30 68862.00 221.51  
## 10: 69.88 20 55642.32 183.82  
## 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  
## 7: Enhanced dedicated support Brandonstad 0 Myanmar  
## 8: Reactive local challenge Port Jefferybury 1 Australia  
## 9: Configurable coherent function West Colin 1 Grenada  
## 10: Mandatory homogeneous architecture Ramirezton 1 Ghana  
## 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  
## 7: 2016-01-28 20:59:32 0  
## 8: 2016-03-07 01:40:15 1  
## 9: 2016-04-18 09:33:42 0  
## 10: 2016-07-11 01:42:51 0

#checking the last 10 rows in our dataset  
(print(tail(df, n=10)))

## Daily Time Spent on Site Age Area Income Daily Internet Usage  
## 1: 35.79 44 33813.08 165.62  
## 2: 38.96 38 36497.22 140.67  
## 3: 69.17 40 66193.81 123.62  
## 4: 64.20 27 66200.96 227.63  
## 5: 43.70 28 63126.96 173.01  
## 6: 72.97 30 71384.57 208.58  
## 7: 51.30 45 67782.17 134.42  
## 8: 51.63 51 42415.72 120.37  
## 9: 55.55 19 41920.79 187.95  
## 10: 45.01 26 29875.80 178.35  
## Ad Topic Line City Male  
## 1: Enterprise-wide tangible model North Katie 1  
## 2: Versatile mission-critical application Mauricefurt 1  
## 3: Extended leadingedge solution New Patrick 0  
## 4: Phased zero tolerance extranet Edwardsmouth 1  
## 5: Front-line bifurcated ability Nicholasland 0  
## 6: Fundamental modular algorithm Duffystad 1  
## 7: Grass-roots cohesive monitoring New Darlene 1  
## 8: Expanded intangible solution South Jessica 1  
## 9: Proactive bandwidth-monitored policy West Steven 0  
## 10: Virtual 5thgeneration emulation Ronniemouth 0  
## Country Timestamp Clicked on Ad  
## 1: Tonga 2016-04-20 13:36:42 1  
## 2: Comoros 2016-07-21 16:02:40 1  
## 3: Montenegro 2016-03-06 11:36:06 1  
## 4: Isle of Man 2016-02-11 23:45:01 0  
## 5: Mayotte 2016-04-04 03:57:48 1  
## 6: Lebanon 2016-02-11 21:49:00 1  
## 7: Bosnia and Herzegovina 2016-04-22 02:07:01 1  
## 8: Mongolia 2016-02-01 17:24:57 1  
## 9: Guatemala 2016-03-24 02:35:54 0  
## 10: Brazil 2016-06-03 21:43:21 1

## Daily Time Spent on Site Age Area Income Daily Internet Usage  
## 1: 35.79 44 33813.08 165.62  
## 2: 38.96 38 36497.22 140.67  
## 3: 69.17 40 66193.81 123.62  
## 4: 64.20 27 66200.96 227.63  
## 5: 43.70 28 63126.96 173.01  
## 6: 72.97 30 71384.57 208.58  
## 7: 51.30 45 67782.17 134.42  
## 8: 51.63 51 42415.72 120.37  
## 9: 55.55 19 41920.79 187.95  
## 10: 45.01 26 29875.80 178.35  
## Ad Topic Line City Male  
## 1: Enterprise-wide tangible model North Katie 1  
## 2: Versatile mission-critical application Mauricefurt 1  
## 3: Extended leadingedge solution New Patrick 0  
## 4: Phased zero tolerance extranet Edwardsmouth 1  
## 5: Front-line bifurcated ability Nicholasland 0  
## 6: Fundamental modular algorithm Duffystad 1  
## 7: Grass-roots cohesive monitoring New Darlene 1  
## 8: Expanded intangible solution South Jessica 1  
## 9: Proactive bandwidth-monitored policy West Steven 0  
## 10: Virtual 5thgeneration emulation Ronniemouth 0  
## Country Timestamp Clicked on Ad  
## 1: Tonga 2016-04-20 13:36:42 1  
## 2: Comoros 2016-07-21 16:02:40 1  
## 3: Montenegro 2016-03-06 11:36:06 1  
## 4: Isle of Man 2016-02-11 23:45:01 0  
## 5: Mayotte 2016-04-04 03:57:48 1  
## 6: Lebanon 2016-02-11 21:49:00 1  
## 7: Bosnia and Herzegovina 2016-04-22 02:07:01 1  
## 8: Mongolia 2016-02-01 17:24:57 1  
## 9: Guatemala 2016-03-24 02:35:54 0  
## 10: Brazil 2016-06-03 21:43:21 1

# Check the structure of the data   
print(str(df))

## Classes 'data.table' and '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 standardization" "Organic bottom-line service-desk" "Triple-buffered reciprocal time-frame" ...  
## $ 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 : POSIXct, format: "2016-03-27 00:53:11" "2016-04-04 01:39:02" ...  
## $ Clicked on Ad : int 0 0 0 0 0 0 0 1 0 0 ...  
## - attr(\*, ".internal.selfref")=<externalptr>   
## NULL

The advitising data is a dataframe/datatable containing 1000 observations and 10 variables. The variables has discrete and countinous values and also it has character values.

#distribution of dataset on gender  
table(df$gender)

## < table of extent 0 >

There are more female that male who use the site

#distribution of dataset on click on ad  
table(df$clicked\_on\_ad)

## < table of extent 0 >

The dataset is well balanced with a ratio of 1:1 of those who click ad and those who did not

## Data Cleaning

### Renaming of columns

# Changing column names to lower and replacing spaces with an underscore for readability  
colnames(df) = tolower(str\_replace\_all(colnames(df), c(' ' = '\_')))  
  
# Checking whether the column names have been renames appriopriately  
print(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"

# The male column should be renamed to gender  
colnames(df)[colnames(df) == 'male'] = 'gender'  
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 gender country  
## 1: Cloned 5thgeneration orchestration Wrightburgh 0 Tunisia  
## 2: Monitored national standardization West Jodi 1 Nauru  
## 3: Organic bottom-line service-desk Davidton 0 San Marino  
## 4: Triple-buffered reciprocal time-frame West Terrifurt 1 Italy  
## 5: Robust logistical utilization South Manuel 0 Iceland  
## 6: Sharable client-driven software Jamieberg 1 Norway  
## timestamp clicked\_on\_ad  
## 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

### Finding missing values

#   
#check missing values in each column  
colSums(is.na(df))

## daily\_time\_spent\_on\_site age area\_income   
## 0 0 0   
## daily\_internet\_usage ad\_topic\_line city   
## 0 0 0   
## gender country timestamp   
## 0 0 0   
## clicked\_on\_ad   
## 0

### Finding Duplicates values

# find the duplicated rows   
#   
anyDuplicated(df)

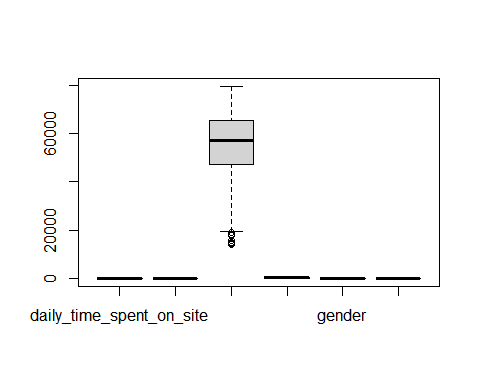
## [1] 0

### Finding Outliers

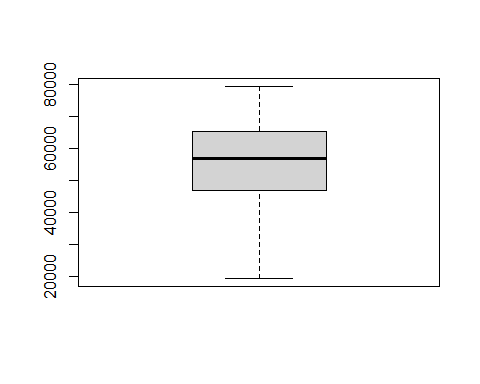
# get numerical columnns  
num <- df %>% select\_if(is.numeric)  
num

## daily\_time\_spent\_on\_site age area\_income daily\_internet\_usage gender  
## 1: 68.95 35 61833.90 256.09 0  
## 2: 80.23 31 68441.85 193.77 1  
## 3: 69.47 26 59785.94 236.50 0  
## 4: 74.15 29 54806.18 245.89 1  
## 5: 68.37 35 73889.99 225.58 0  
## ---   
## 996: 72.97 30 71384.57 208.58 1  
## 997: 51.30 45 67782.17 134.42 1  
## 998: 51.63 51 42415.72 120.37 1  
## 999: 55.55 19 41920.79 187.95 0  
## 1000: 45.01 26 29875.80 178.35 0  
## clicked\_on\_ad  
## 1: 0  
## 2: 0  
## 3: 0  
## 4: 0  
## 5: 0  
## ---   
## 996: 1  
## 997: 1  
## 998: 1  
## 999: 0  
## 1000: 1

#check for outliers in the numerical columns  
boxplot(num)

 From the results we can conclude that there are outliers in Area Income columns we will not delete them because we will get insights from them

non\_outliers <- 47032 - 1.5 \* IQR(df$area\_income)   
df$area\_income[df$area\_income < non\_outliers]<- non\_outliers  
  
boxplot(df$area\_income)



# Exploratory Data Analysis

## Univariate Analysis

# checking for the dataset statistical summary  
(summary(num))

## 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   
## gender 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

### Mean

print("Age")

## [1] "Age"

mean(replace(df$age, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 40.334

print("Area Income")

## [1] "Area Income"

mean(replace(df$area\_income, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 48665.06

print("Daily internet usage")

## [1] "Daily internet usage"

mean(replace(df$daily\_internet\_usage, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 145.4865

print("Daily time spent on site")

## [1] "Daily time spent on site"

mean(replace(df$daily\_time\_spent\_on\_site, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 53.14578

### range

print("Age")

## [1] "Age"

range(replace(df$age, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 19 61

print("Area Income")

## [1] "Area Income"

range(replace(df$area\_income, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 19373.75 78520.99

print("Daily internet usage")

## [1] "Daily internet usage"

range(replace(df$daily\_internet\_usage, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 104.78 269.96

print("Daily time spent on site")

## [1] "Daily time spent on site"

range(replace(df$daily\_time\_spent\_on\_site, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 32.60 91.37

### Interquartile range

# Finding the Quantiles for numerical columns  
# Quantiles of daily time on site  
print("Quantiles of daily time on site")

## [1] "Quantiles of daily time on site"

daily\_time\_quantiles <-quantile(df$daily\_time\_on\_site)  
daily\_time\_quantiles

## 0% 25% 50% 75% 100%   
## NA NA NA NA NA

# Quantiles of age  
print("Quantiles of age")

## [1] "Quantiles of age"

age\_quantiles <-quantile(df$age)  
age\_quantiles

## 0% 25% 50% 75% 100%   
## 19 29 35 42 61

# Quantiles of area income  
print("Quantiles of area income")

## [1] "Quantiles of area income"

area\_income\_quantiles <-quantile(df$area\_income)  
area\_income\_quantiles

## 0% 25% 50% 75% 100%   
## 19373.75 47031.80 57012.30 65470.64 79484.80

# Quantiles of daily internet usage  
print("Quantiles of daily internet usage")

## [1] "Quantiles of daily internet usage"

daily\_internet\_usage\_quantiles <-quantile(df$daily\_internet\_usage)  
daily\_internet\_usage\_quantiles

## 0% 25% 50% 75% 100%   
## 104.7800 138.8300 183.1300 218.7925 269.9600

### Standard deviation

print("Age")

## [1] "Age"

sd(replace(df$age, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 8.856598

print("Area Income")

## [1] "Area Income"

sd(replace(df$area\_income, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 14003.2

print("Daily internet usage")

## [1] "Daily internet usage"

sd(replace(df$daily\_internet\_usage, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 30.02583

print("Daily time spent on site")

## [1] "Daily time spent on site"

sd(replace(df$daily\_time\_spent\_on\_site, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 12.82209

### Variance

print("Age")

## [1] "Age"

var(replace(df$age, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 78.43932

print("Area Income")

## [1] "Area Income"

var(replace(df$area\_income, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 196089721

print("Daily internet usage")

## [1] "Daily internet usage"

var(replace(df$daily\_internet\_usage, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 901.5502

print("Daily time spent on site")

## [1] "Daily time spent on site"

var(replace(df$daily\_time\_spent\_on\_site, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 164.406

### Skewness

#skewness of Area Income for those who click ad  
library(moments)  
print("Age")

## [1] "Age"

skewness(replace(df$age, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 0.02515906

print("Area Income")

## [1] "Area Income"

skewness(replace(df$area\_income, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] -0.1307072

print("Daily internet usage")

## [1] "Daily internet usage"

skewness(replace(df$daily\_internet\_usage, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 1.236424

print("Daily time spent on site")

## [1] "Daily time spent on site"

skewness(replace(df$daily\_time\_spent\_on\_site, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 0.5337215

All the variables has positive value this implies that the distribution of the data is slightly skewed to the right or positive skewed. accept for area income which negative which means it is skewed to left

### Kurtosis

print("Age")

## [1] "Age"

kurtosis(replace(df$age, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 2.303378

print("Area Income")

## [1] "Area Income"

kurtosis(replace(df$area\_income, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 2.286888

print("Daily internet usage")

## [1] "Daily internet usage"

kurtosis(replace(df$daily\_internet\_usage, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 4.816968

print("Daily time spent on site")

## [1] "Daily time spent on site"

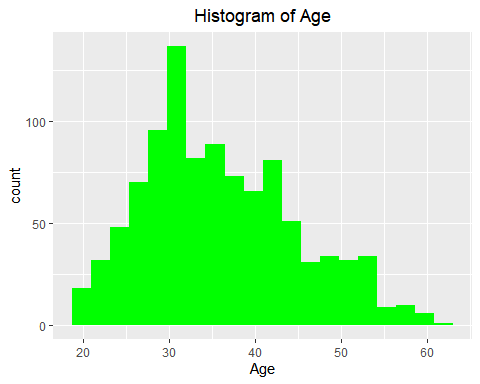
kurtosis(replace(df$daily\_time\_spent\_on\_site, df$clicked\_on\_ad==0, NA), na.rm = TRUE)

## [1] 2.561506

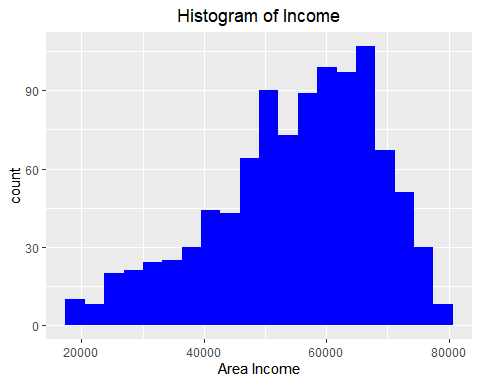
The distribution of the data is platykurtic, since the computed values for all the varibles are less than 3 accept variable daily internet usage.

### Histogram

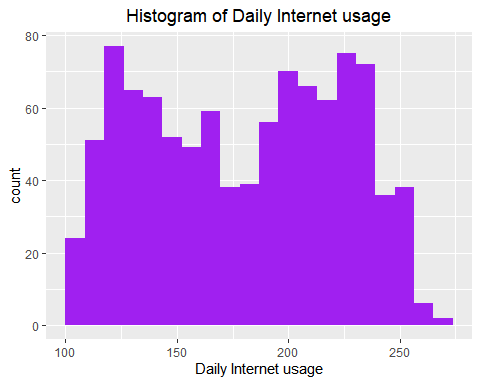
# Histogram of Age  
ggplot(data = df, mapping = aes(x = age)) +  
 geom\_histogram(bins = 20, fill = "green") +  
 labs(x = "Age") +  
 ggtitle("Histogram of Age") +  
 theme(plot.title = element\_text(hjust = 0.5))

 It is skewed to left. Age between 25 and 45 visit the site mostly

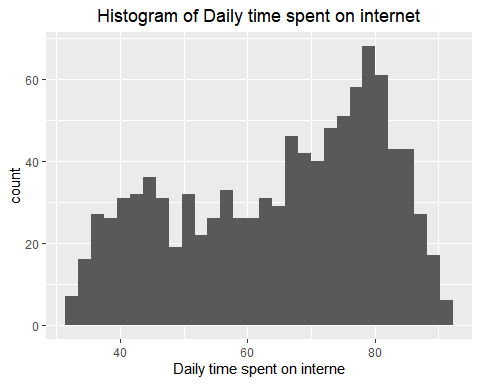
# Histogram of Income  
ggplot(data = df, mapping = aes(x = area\_income)) +  
 geom\_histogram(bins = 20, fill = "blue") +  
 labs(x = "Area Income") +  
 ggtitle("Histogram of Income") +  
 theme(plot.title = element\_text(hjust = 0.5))

 It is skewed to right, people whose income ranges between 45000 and 75000 mostly visit the site.

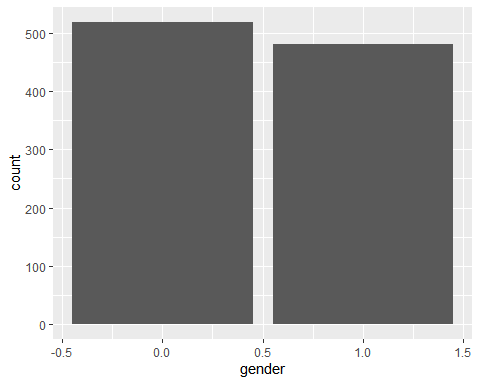
# Histogram of Daily Internet usage  
ggplot(data = df, mapping = aes(x = daily\_internet\_usage)) +  
 geom\_histogram(bins = 20,fill="purple") +  
 labs(x = "Daily Internet usage") +  
 ggtitle("Histogram of Daily Internet usage") +  
 theme(plot.title = element\_text(hjust = 0.5))



# Histogram of Daily time spent on internet  
ggplot(data = df, mapping = aes(x = daily\_time\_spent\_on\_site)) +  
 geom\_histogram(bins = 30) +  
 labs(x = "Daily time spent on interne") +  
 ggtitle("Histogram of Daily time spent on internet") +  
 theme(plot.title = element\_text(hjust = 0.5))

 Most of the people who are above 60 visits the site mostly

#Which gender is mainly active on the blog?  
ggplot(data = df) +  
 geom\_bar(mapping = aes(x = gender))



#Assuming that if male = 1 then we can conclude that more females  
# frequennt the blog more as compared to males

Female visit the site mostly than male

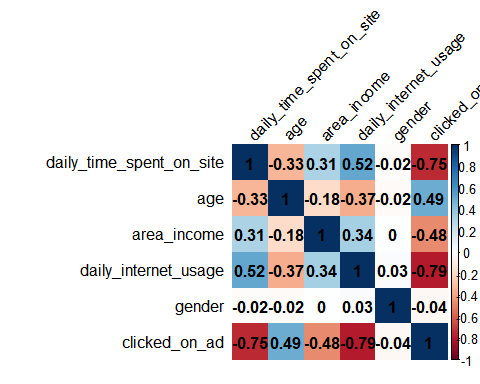
## Bivariate and Multivariate Analysis.

### Correlation

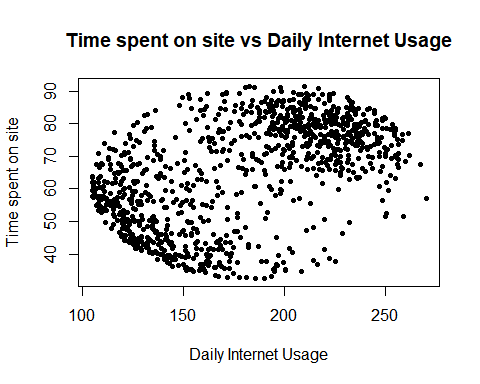
library(corrplot)

## corrplot 0.84 loaded

#Get the correlation matrix  
res = cor(num)  
#Plotting a correlation plot  
  
corrplot(res, method="color",addCoef.col = "black",   
 tl.col="black", tl.srt=45)

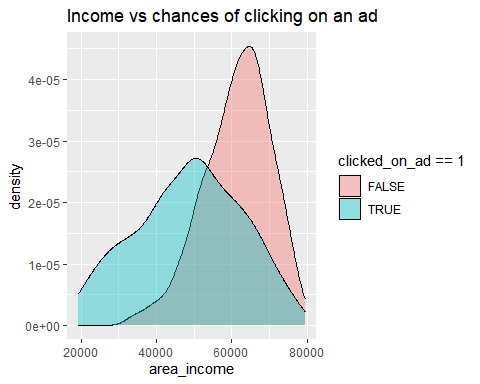
 There are a bit of correlation between Daily time spent on site and Daily internet usage

x <- df$daily\_internet\_usage  
y <- df$daily\_time\_spent\_on\_site  
# Plot with main and axis titles  
# Change point shape (pch = 19) and remove frame.  
plot(x, y, main = "Time spent on site vs Daily Internet Usage",  
 xlab = "Daily Internet Usage", ylab = "Time spent on site",  
 pch = 20)

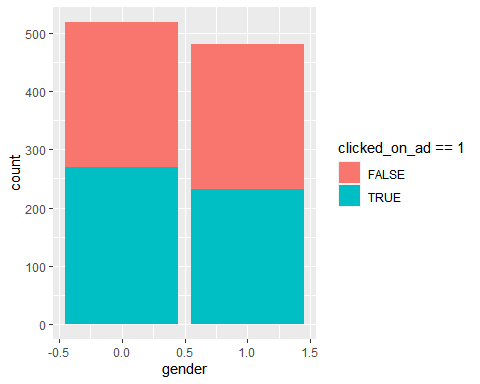


The plot shows that there is no correlation between the two columns. But we can see that people who spend less time on site use less internet. Also, most of the people who use alot of internet per day seem to spend most of there time on the site.

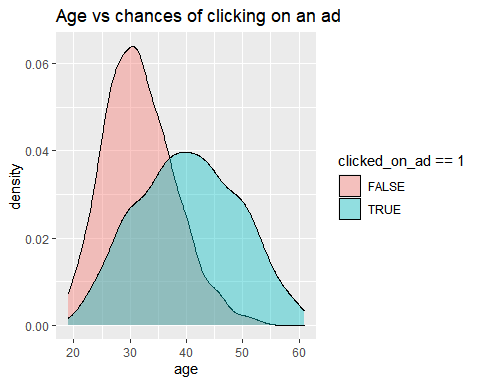
#Income class and it's relationship to clicking an ad  
ggplot(df,   
 aes(x = area\_income,   
 fill = clicked\_on\_ad==1)) +  
 geom\_density(alpha = 0.4) +  
 labs(title = "Income vs chances of clicking on an ad")

 People from all range of income click on ads as compared to those who do not click ad. should focus on people whose income range on 20000 to 50000 are the ones are likely click ad

#Who is likely to click on an ad, female or male?  
# stacked bar chart  
library(ggplot2)  
ggplot(df,   
 aes(x = gender,   
 fill = clicked\_on\_ad==1)) +   
 geom\_bar(position = "stack")

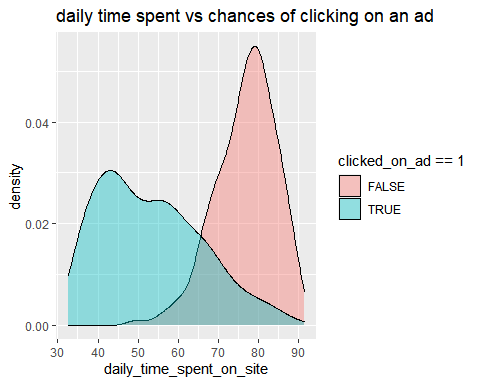
 There are more female who click on ad than male

#Age and it's relationship to clicking an ad  
ggplot(df,   
 aes(x = age,   
 fill = clicked\_on\_ad==1)) +  
 geom\_density(alpha = 0.4) +  
 labs(title = "Age vs chances of clicking on an ad")



People who click on ad are disriputed across the age. Those who are in the bracket of 30 to 50 are the ones who click on add mostly

#Time on and it's relationship to clicking an ad  
ggplot(df,   
 aes(x = daily\_time\_spent\_on\_site,   
 fill = clicked\_on\_ad==1)) +  
 geom\_density(alpha = 0.4) +  
 labs(title = "daily time spent vs chances of clicking on an ad")



People who spend less time on site click on ad unlike those who spend longer time on site

# Conclusion

1. Most of the people who use alot of internet per day seem to spend most of there time on the site.She should target those who spend alot of internet.
2. Those who are in the age bracket of 30 to 50 are the ones who click on ad mostly so she should focus on people who are in age bracket named.
3. Famele who click on ad are more than male so she should forcus on them, without neglecting male because no big difference between them interms of clicking on ad
4. Those who click on ad is distributed across income but those who are in the range of 20000 to 50000 click on ad mostly, so she should consider them.
5. Those who spend less than 70mins on site mostly click on ad