**1. Defining the question**

## 1.1 Specifying the data analytic objective

Predict which individuals are most likely to click on ads from a cryptography course website

## 1.2 Defining the metric of success

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

## 1.3 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.

## 1.4 Recording the Experimental Design

1. Loading the data
2. Checking the data
3. Tidying the data
4. Univariate Analysis
5. Bivariate Analysis
6. Challenging the solution
7. Recommendations
8. Follow up questions

# 2. Loading the data set

library(data.table)  
advert <- fread('http://bit.ly/IPAdvertisingData')  
advert

## 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  
## ---   
## 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  
## 1: Cloned 5thgeneration orchestration Wrightburgh 0  
## 2: Monitored national standardization West Jodi 1  
## 3: Organic bottom-line service-desk Davidton 0  
## 4: Triple-buffered reciprocal time-frame West Terrifurt 1  
## 5: Robust logistical utilization South Manuel 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  
## 1: Tunisia 2016-03-27 00:53:11 0  
## 2: Nauru 2016-04-04 01:39:02 0  
## 3: San Marino 2016-03-13 20:35:42 0  
## 4: Italy 2016-01-10 02:31:19 0  
## 5: Iceland 2016-06-03 03:36:18 0  
## ---   
## 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

### Checking the data summary

summary(advert)

## 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   
## Ad Topic Line City Male Country   
## Length:1000 Length:1000 Min. :0.000 Length:1000   
## Class :character Class :character 1st Qu.:0.000 Class :character   
## Mode :character Mode :character Median :0.000 Mode :character   
## Mean :0.481   
## 3rd Qu.:1.000   
## Max. :1.000   
## Timestamp Clicked on Ad  
## Min. :2016-01-01 02:52:10 Min. :0.0   
## 1st Qu.:2016-02-18 02:55:42 1st Qu.:0.0   
## Median :2016-04-07 17:27:29 Median :0.5   
## Mean :2016-04-10 10:34:06 Mean :0.5   
## 3rd Qu.:2016-05-31 03:18:14 3rd Qu.:1.0   
## Max. :2016-07-24 00:22:16 Max. :1.0

From the data summary we get the measures of central tendency (median, mean, mode and quantile)

### Checking the top and bottom columns

tail(advert)

## Daily Time Spent on Site Age Area Income Daily Internet Usage  
## 1: 43.70 28 63126.96 173.01  
## 2: 72.97 30 71384.57 208.58  
## 3: 51.30 45 67782.17 134.42  
## 4: 51.63 51 42415.72 120.37  
## 5: 55.55 19 41920.79 187.95  
## 6: 45.01 26 29875.80 178.35  
## Ad Topic Line City Male  
## 1: Front-line bifurcated ability Nicholasland 0  
## 2: Fundamental modular algorithm Duffystad 1  
## 3: Grass-roots cohesive monitoring New Darlene 1  
## 4: Expanded intangible solution South Jessica 1  
## 5: Proactive bandwidth-monitored policy West Steven 0  
## 6: Virtual 5thgeneration emulation Ronniemouth 0  
## Country Timestamp Clicked on Ad  
## 1: Mayotte 2016-04-04 03:57:48 1  
## 2: Lebanon 2016-02-11 21:49:00 1  
## 3: Bosnia and Herzegovina 2016-04-22 02:07:01 1  
## 4: Mongolia 2016-02-01 17:24:57 1  
## 5: Guatemala 2016-03-24 02:35:54 0  
## 6: Brazil 2016-06-03 21:43:21 1

head(advert)

## 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

class(advert)

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

### Structure of the dataset

str(advert)

## 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>

#3. Cleaning the dataset

##3.1 Finding missing values

colSums(is.na(advert))

## Daily Time Spent on Site Age Area Income   
## 0 0 0   
## Daily Internet Usage Ad Topic Line City   
## 0 0 0   
## Male Country Timestamp   
## 0 0 0   
## Clicked on Ad   
## 0

No missing data was found

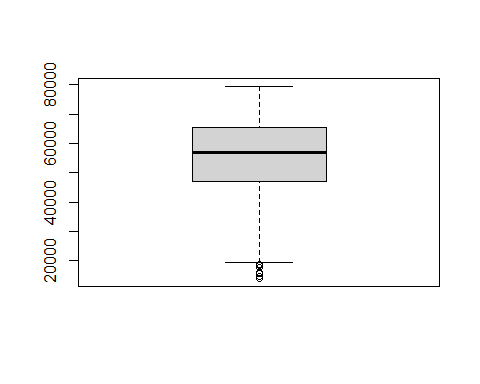
## 3.2 Checking for duplicates

sum(duplicated(advert))

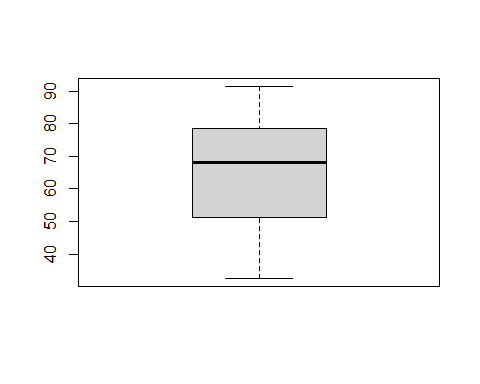
## [1] 0

## 3.3 Checking for outliers

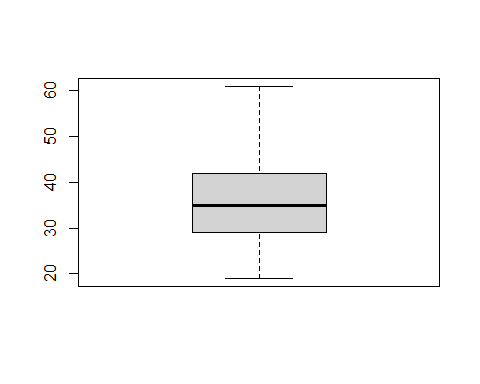
# Area Income  
boxplot(advert$`Area Income`)



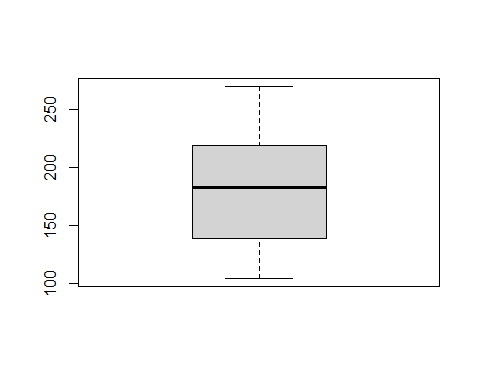
# Time spent on site  
boxplot(advert$`Daily Time Spent on Site`)



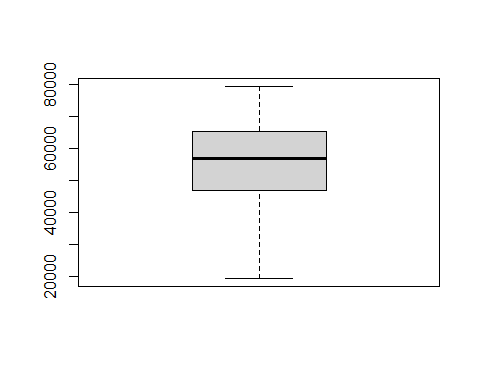
# Age  
boxplot(advert$Age)



# Daily internet usage  
boxplot(advert$`Daily Internet Usage`)

 ##3.4 Removing outliers

outlier <- 47032 - 1.5 \* IQR(advert$`Area Income`)   
advert$`Area Income`[advert$`Area Income` < outlier]<- outlier  
  
boxplot(advert$`Area Income`)



We remove outliers by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers

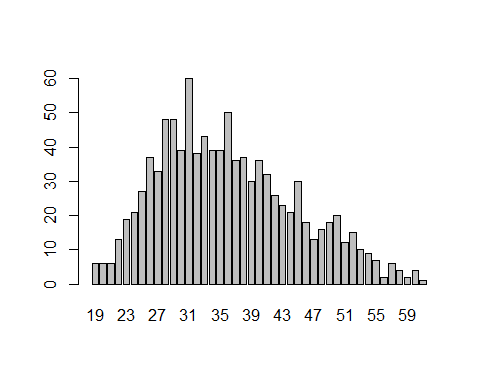
# 4. Exploratory Data Analysis

## 4.1 Univariate Analysis

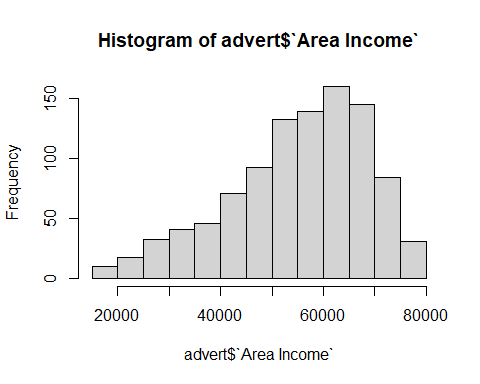
# Age Frequency  
# fetching the age  
age <- advert$Age  
age\_freq <- table(age)  
age\_freq

## age  
## 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44   
## 6 6 6 13 19 21 27 37 33 48 48 39 60 38 43 39 39 50 36 37 30 36 32 26 23 21   
## 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61   
## 30 18 13 16 18 20 12 15 10 9 7 2 6 4 2 4 1

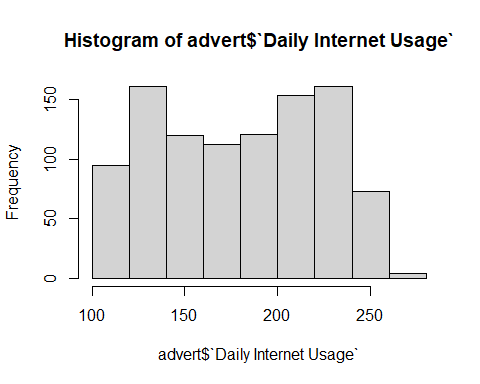
# Creating a bar graph of age  
barplot(age\_freq)



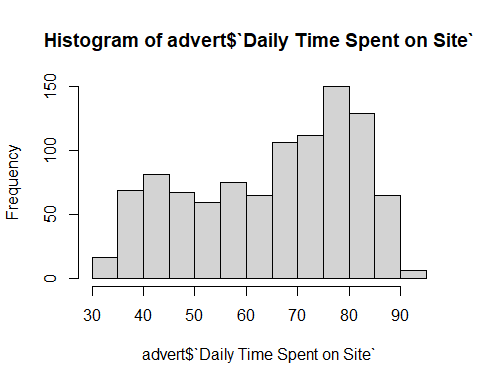
# Histogram for area income  
hist(advert$`Area Income`)



# Histogram for area income  
hist(advert$`Daily Internet Usage`)



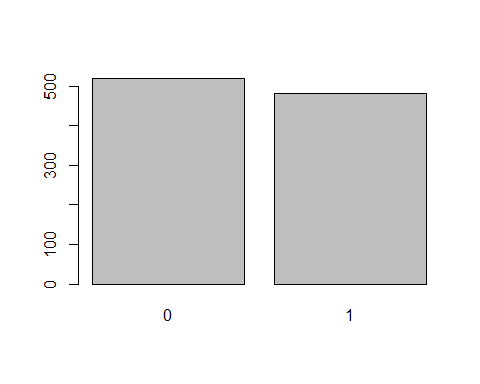
# Histogram for Daily Time  
hist(advert$`Daily Time Spent on Site`)



# Gender Frequency  
# fetching the male column  
male\_female <- advert$Male  
gender\_freq <- table(male\_female)  
gender\_freq

## male\_female  
## 0 1   
## 519 481

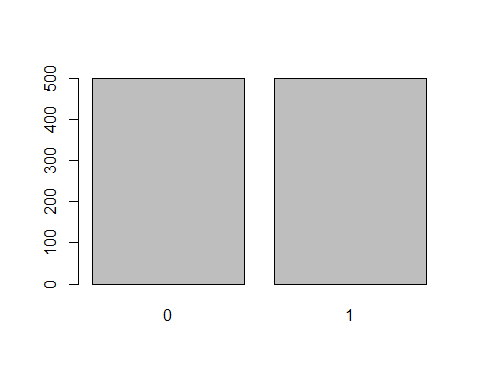
# Creating a bar graph of age  
barplot(gender\_freq)



# Clicked on Ad Frequency  
# fetching the Clicked on Ad  
Clicked.on.Ad <- advert$`Clicked on Ad`  
clicked\_freq <- table(Clicked.on.Ad)  
clicked\_freq

## Clicked.on.Ad  
## 0 1   
## 500 500

# Creating a bar graph of clicked on age  
barplot(clicked\_freq)



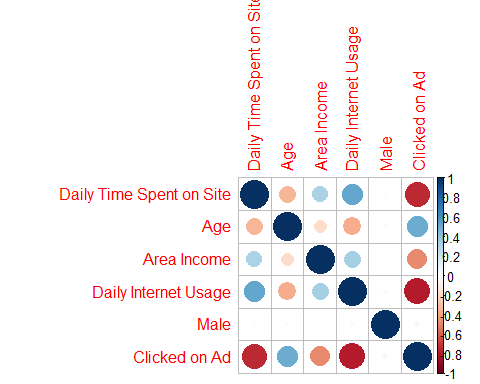
## 4.2 Bivariate analysis

Here we check for correlation between the different columns and the target variable Clicked on ad

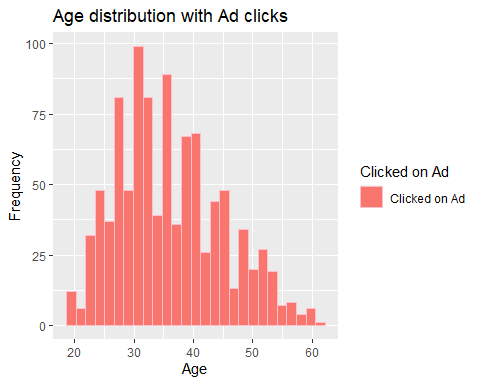
library(corrplot)

## corrplot 0.84 loaded

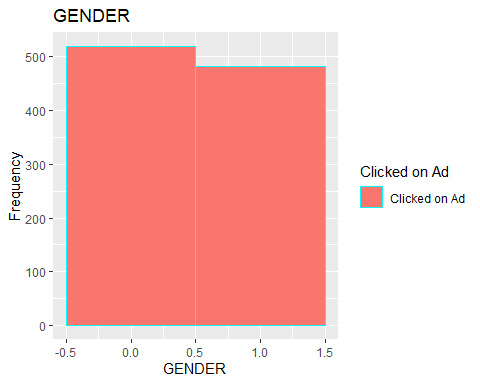
advert\_num <- Filter(is.numeric, advert)  
corrplot(cor(advert\_num))



library(ggplot2)  
  
ggplot(data = advert, aes(x = Age, fill = 'Clicked on Ad' ))+  
 geom\_histogram(bins = 27, color = 'pink') +   
 labs(title = 'Age distribution with Ad clicks', x = 'Age', y = 'Frequency', fill = 'Clicked on Ad')



ggplot(data = advert, aes(x = Male, fill = 'Clicked on Ad'))+  
 geom\_histogram(bins = 2, color = 'cyan') +   
 labs(title = 'GENDER', x = 'GENDER', y = 'Frequency', fill = 'Clicked on Ad') +  
 scale\_color\_brewer(palette = 'Set1')



# Conclusion

The ages between 26 and 42 record the highest frequency of ad clicks on the site and also the highest amount of time spent on the internet.