# 1. Defining the Question

### a) Specifying the Data Analytic Question

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

### b) Defining the Metric for Success

For this sttudy, 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) 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

# Reading the data

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

## 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 summary  
summary(ads)

## 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

# checking the dimension/shape of the data  
dim(ads) # 1000 rows and 10 columns

## [1] 1000 10

#Change datatypes  
ads$Male <- as.factor(ads$Male)  
#   
#Checking datatypes  
sapply(ads, class)

## $`Daily Time Spent on Site`  
## [1] "numeric"  
##   
## $Age  
## [1] "integer"  
##   
## $`Area Income`  
## [1] "numeric"  
##   
## $`Daily Internet Usage`  
## [1] "numeric"  
##   
## $`Ad Topic Line`  
## [1] "character"  
##   
## $City  
## [1] "character"  
##   
## $Male  
## [1] "factor"  
##   
## $Country  
## [1] "character"  
##   
## $Timestamp  
## [1] "POSIXct" "POSIXt"   
##   
## $`Clicked on Ad`  
## [1] "integer"

# Data cleaning

## Missing values

# checking for missing values  
sum(is.na(ads))# there are no missing values in the data

## [1] 0

# displaying all rows from the dataset which don't contain any missing values  
na.omit(ads)

## 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

## Duplicates

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.1.1 v dplyr 1.0.6  
## 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::between() masks data.table::between()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::first() masks data.table::first()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::last() masks data.table::last()  
## x purrr::transpose() masks data.table::transpose()

# Checking the number of duplicates  
duplicated\_rows <- ads[duplicated(ads),]  
duplicated\_rows # there are no duplicates in the data

## Empty data.table (0 rows and 10 cols): Daily Time Spent on Site,Age,Area Income,Daily Internet Usage,Ad Topic Line,City...

# Data checking

# Datatypes  
sapply(ads,class)

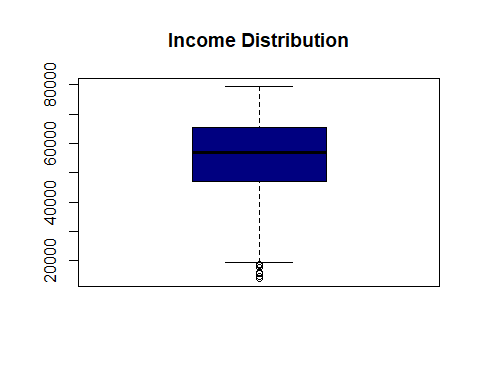
## $`Daily Time Spent on Site`  
## [1] "numeric"  
##   
## $Age  
## [1] "integer"  
##   
## $`Area Income`  
## [1] "numeric"  
##   
## $`Daily Internet Usage`  
## [1] "numeric"  
##   
## $`Ad Topic Line`  
## [1] "character"  
##   
## $City  
## [1] "character"  
##   
## $Male  
## [1] "factor"  
##   
## $Country  
## [1] "character"  
##   
## $Timestamp  
## [1] "POSIXct" "POSIXt"   
##   
## $`Clicked on Ad`  
## [1] "integer"

# Checking for outliers

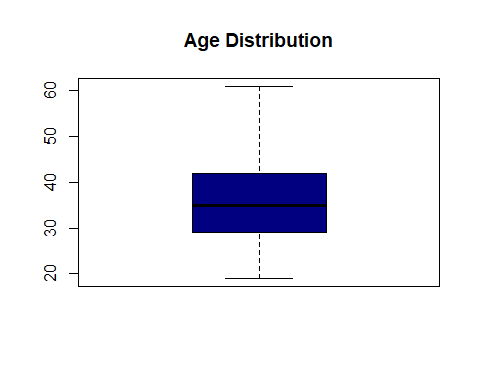
#Checking for outliers  
#First we select numeric columns excluding male and clicked.on.ad since they are binary column  
  
ads\_df <- subset(ads, select = -c(`Ad Topic Line`,City, Male, Country, Timestamp, `Clicked on Ad`))  
head(ads\_df)

## 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

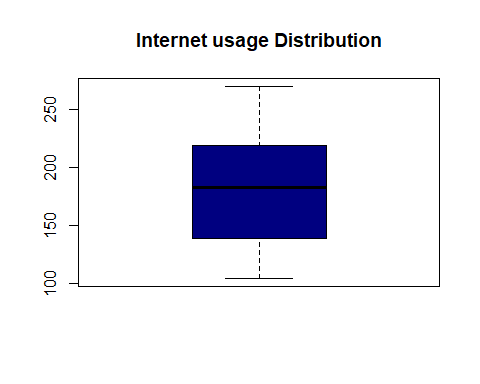
# Then use the function boxplot.stats which lists the outliers in the vectors  
# ---  
#  
options(repr.plot.width = 14, repr.plot.height = 14)  
boxplot(ads$`Area Income`,  
 col = "navyblue")  
# Add a title  
title("Income Distribution")



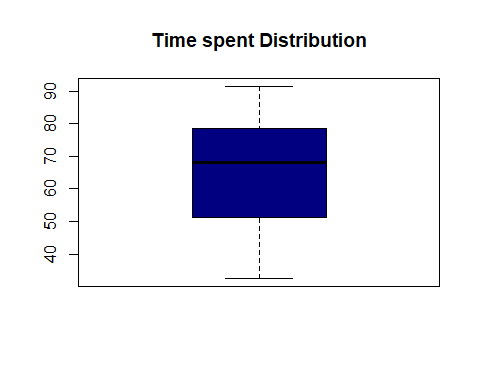
# Then use the function boxplot.stats which lists the outliers in the vectors  
# ---  
#  
options(repr.plot.width = 14, repr.plot.height = 10)  
boxplot(ads$`Age`,  
 col = "navyblue")  
# Add a title  
title("Age Distribution")



# Then use the function boxplot.stats which lists the outliers in the vectors  
# ---  
#  
options(repr.plot.width = 14, repr.plot.height = 14)  
boxplot(ads$`Daily Internet Usage`,  
 col = "navyblue")  
# Add a title  
title("Internet usage Distribution")



# Then use the function boxplot.stats which lists the outliers in the vectors  
# ---  
#  
options(repr.plot.width = 10, repr.plot.height = 10)  
boxplot(ads$`Daily Time Spent on Site`,  
 col = "navyblue")  
# Add a title  
title("Time spent Distribution")



class(ads)

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

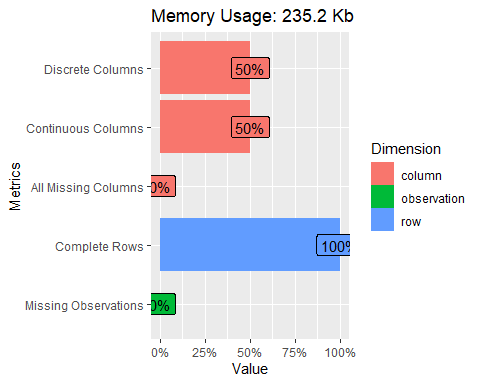
str(ads)

## 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 : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 2 2 ...  
## $ 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>

library(tidyverse)  
library(DataExplorer)  
glimpse(ads)

## Rows: 1,000  
## Columns: 10  
## $ `Daily Time Spent on Site` <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, 8~  
## $ Age <int> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49,~  
## $ `Area Income` <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 738~  
## $ `Daily Internet Usage` <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 226~  
## $ `Ad Topic Line` <chr> "Cloned 5thgeneration orchestration", "Moni~  
## $ City <chr> "Wrightburgh", "West Jodi", "Davidton", "We~  
## $ Male <fct> 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0~  
## $ Country <chr> "Tunisia", "Nauru", "San Marino", "Italy", ~  
## $ Timestamp <dttm> 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, 1, 0, 1, 0, 1~

options(repr.plot.width = 10, repr.plot.height = 10)  
plot\_intro(ads)



# Exploratory Data Analysis

## Univariate analysis

##### Measures of central tendency

# descriptive statistics  
# these summaries will provide us with the measures of central tendencies of the numerical columns  
summary(ads)

## 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 0:519 Length:1000   
## Class :character Class :character 1:481 Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## 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

###### Age

# Finding the mode of age  
getmode <- function(v) {  
 uniqv <- unique(v)  
 uniqv[which.max(tabulate(match(v, uniqv)))]  
}  
  
ads.Age.mode <- getmode(ads$Age)  
ads.Age.mode

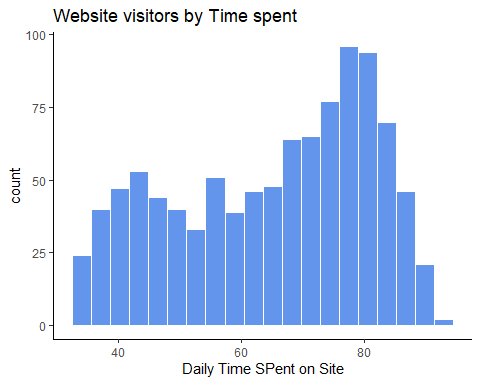
## [1] 31

# Finding the median of age  
ads.Age.median <- median(ads$Age)  
ads.Age.median

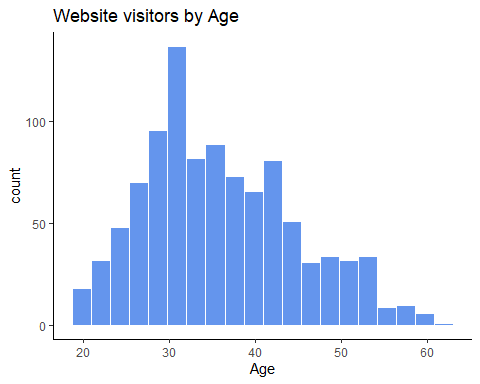
## [1] 35

## Univariate graphical

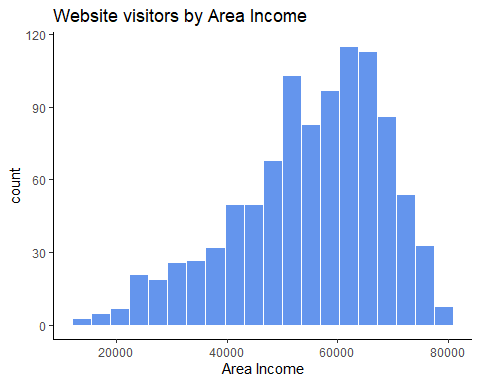
ggplot(ads, aes(x = `Daily Time Spent on Site`)) +  
 geom\_histogram(fill = "cornflowerblue",   
 color = "white",bins = 20) +   
 theme\_classic() + # use a minimal theme  
 labs(title="Website visitors by Time spent",  
 x = "Daily Time SPent on Site")



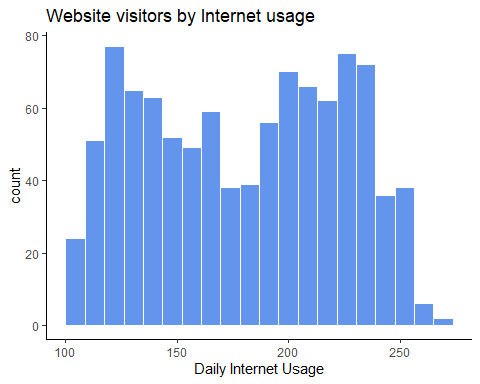
ggplot(ads, aes(x = Age)) +  
 geom\_histogram(fill = "cornflowerblue",   
 color = "white",bins = 20) +   
 theme\_classic() + # use a minimal theme  
 labs(title="Website visitors by Age",  
 x = "Age")



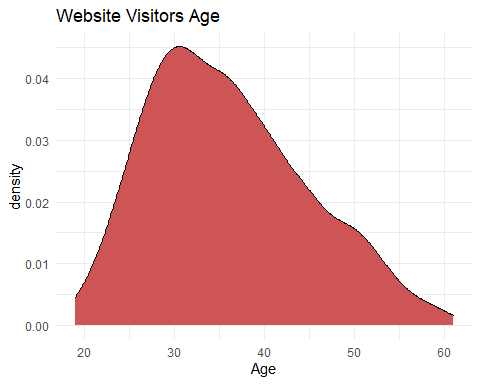
library(ggplot2) # for visualization  
ggplot(ads, aes(x = `Area Income`)) +  
 geom\_histogram(fill = "cornflowerblue",   
 color = "white",bins = 20) +   
 theme\_classic() + # use a minimal theme  
 labs(title="Website visitors by Area Income",  
 x = "Area Income")



library(ggplot2) # for visualization  
ggplot(ads, aes(x = `Daily Internet Usage`)) +  
 geom\_histogram(fill = "cornflowerblue",   
 color = "white",bins = 20) +   
 theme\_classic() + # use a minimal theme  
 labs(title="Website visitors by Internet usage",  
 x = "Daily Internet Usage")



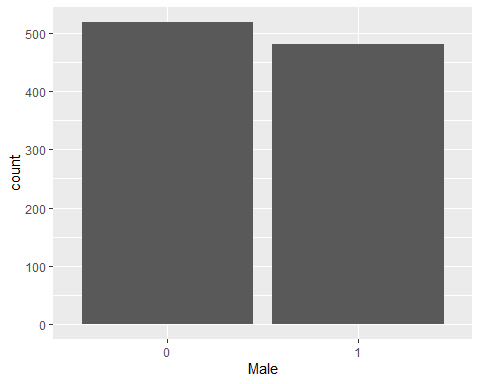
library(ggplot2) # for visualization  
ggplot(ads, aes(x = Age)) +  
 geom\_density(fill = "indianred3") +  
 theme\_minimal() + # use a minimal theme  
 labs(title = "Website Visitors Age")



Age: Most people who visit the blog are between 25 and 40 years, data is skewed to the right of the mean. Graph doesn’t show a sharp peak Income: Data on income is mostly skewed to the right of the 55,00 mean Time spent on site: There are lots of variations on how much time people spend on the site. A good number does spend between 65 and 85 time on the site.

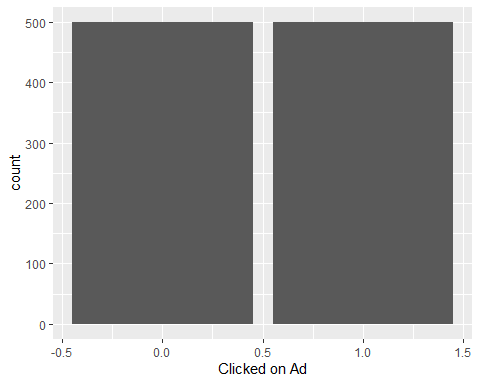
###### Categorical data

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



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

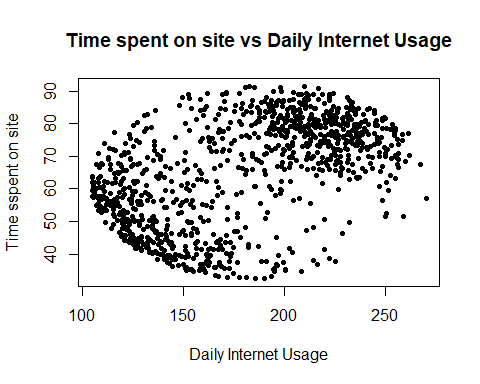
#Do most people clickon ads or not?  
ggplot(data = ads) +  
 geom\_bar(mapping = aes(x = `Clicked on Ad`))



#There is a balance between the people who click on ads and those who don't  
#We can coclude that half of the people who visit the website will probably click on an ad

## Bivariate analysis and Mulitivariate analysis

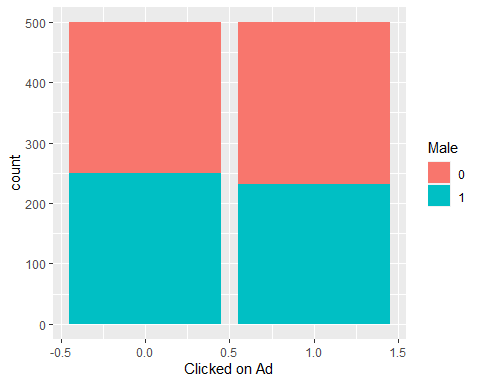
x <- ads$`Daily Internet Usage`  
y <- ads$`Daily Time Spent on Site`  
# Plot with main and axis titles  
# Change point shape (pch = 19) and remove frame.  
plot(x, y, main = "Time spent on site vs Daily Internet Usage",  
 xlab = "Daily Internet Usage", ylab = "Time sspent on site",  
 pch = 20)



The points are all over but our data points are not highly correlated which explains this. 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 a considerable amount of time on the site. This is a good sign for the business

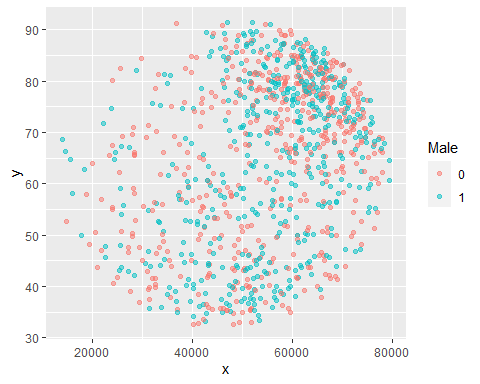
##### Explore which individuals are most likely to click an ad

#Who is likely to click on an ad, female or male?  
library(ggplot2)  
  
# stacked bar chart  
ggplot(ads,   
 aes(x = `Clicked on Ad`,   
 fill = Male)) +   
 geom\_bar(position = "stack")



From above graph, we can see that female(where male=0), is slightly more likely to click on an ad that male. But also from earlier analysis, it is important to note that females are slightly more frequent on the blog than male.

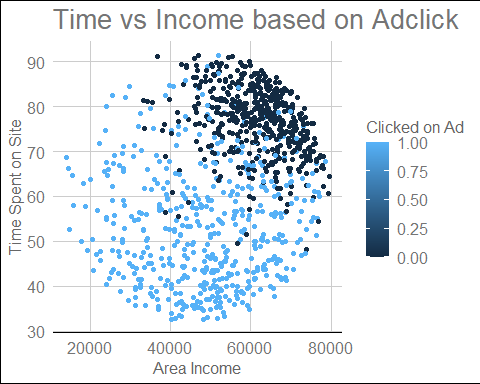
x <- ads$`Area Income`  
y <- ads$`Daily Time Spent on Site`  
ggplot(ads,aes(x,y,color=`Male`))+geom\_point(alpha=0.5)



# finding the correlation  
#rquery.cormat(df1, type="full")  
res <- cor(ads\_df)  
round(res, 2)

## Daily Time Spent on Site Age Area Income  
## Daily Time Spent on Site 1.00 -0.33 0.31  
## Age -0.33 1.00 -0.18  
## Area Income 0.31 -0.18 1.00  
## Daily Internet Usage 0.52 -0.37 0.34  
## Daily Internet Usage  
## Daily Time Spent on Site 0.52  
## Age -0.37  
## Area Income 0.34  
## Daily Internet Usage 1.00

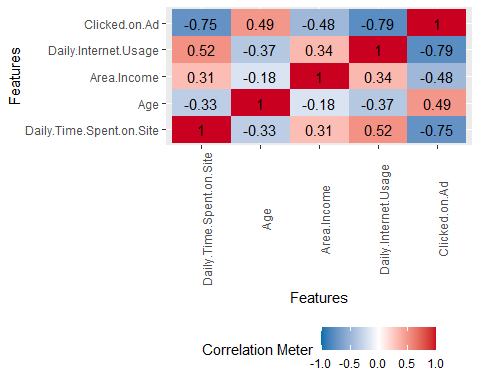
# Scatter Plot for Time vs Income based on adclicks  
  
library(ggthemes)  
library(ggplot2)  
options(repr.plot.width = 10, repr.plot.height = 10)  
ggplot(data = ads, mapping = aes(`Area Income` , `Daily Time Spent on Site`)) +  
 geom\_point(alpha = 10, aes(color = `Clicked on Ad`)) + theme\_gdocs() +  
 labs(title = "Time vs Income based on Adclick",  
 x = "Area Income", y = "Time Spent on Site")



library(corrplot)

## corrplot 0.88 loaded

library(DataExplorer)  
options(repr.plot.width = 20, repr.plot.height = 20)  
plot\_correlation(ads, type = 'c',cor\_args = list( 'use' = 'complete.obs'))



### 6. Conclusion

People who have a daily internet usage of less than 175 are more likely to click on an ad People who spend less than 70mins on the site are likely to click on ad People above 40 are more likely to click on an ad People with an income of less than 60000 are most likely to click on an ad The female gender is likely to click on ad

### 7. Challenging the solution

It would be great to do some hypothesis testing on the conclusions made from Exploratory Data Analysis, this way we could ascertain the chances of specific person clicking on an ad or not. Also, it would be necessary to create a predictive model and perform some feature importance selection to choose which variables are most important to use when deciding who will click on an ad or not when using the website.