Week13\_AdvertisingAnalysis

Stacy Mwende

3/6/2020

# 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.

# Defining the question

Identify which individuals are most likely to click on her ads given different variables for consideration.

# Metric for success

Identify the most appropriate variables to consider while putting up the ad to help achieve the goal which is enrollment to the cryptography course.

# Experimental Design

To help answer the question inplace, different stages were involved as stated below: Business understanding Data understanding Analysis and Modelling Conclusion and Recommendation

# Appropriateness of data

Based on the various variables provided for analysis, RESEARCH ON THIS

# Loading the Dataset

## Get the work directory for the dataset

directory <- "C:/Users/comp5/Downloads/R\_Program"  
data <- file.path(directory,"advertising.csv")  
advert <- read.csv(data) # Reading the dataset  
head(advert) # A print of the first 6 rows

## 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

# Data understanding

**Getting the number of rows and columns**

print(nrow(advert))

## [1] 1000

print(ncol(advert))

## [1] 10

The daaset has 1000 entries and 10 variables

**Getting the datatypes**

sapply(advert, class)

## Daily.Time.Spent.on.Site Age Area.Income   
## "numeric" "integer" "numeric"   
## Daily.Internet.Usage Ad.Topic.Line City   
## "numeric" "factor" "factor"   
## Male Country Timestamp   
## "integer" "factor" "factor"   
## Clicked.on.Ad   
## "integer"

The Timestamp datatype will be converted to the appropriate datatype in the datacleaning section.

# Data Cleaning

**Check for null values**

sum(is.na(advert))

## [1] 0

There is no missing data in the provided dataset

**Check for duplicates**

anyDuplicated(advert)

## [1] 0

There are no duplicates in the dataset

**Correcting wrongly labelled column names** Rename the Male column name to Gender

names(advert)[names(advert) == "Male"] <- "Gender"  
print(colnames(advert))

## [1] "Daily.Time.Spent.on.Site" "Age"   
## [3] "Area.Income" "Daily.Internet.Usage"   
## [5] "Ad.Topic.Line" "City"   
## [7] "Gender" "Country"   
## [9] "Timestamp" "Clicked.on.Ad"

The male column has been corrected to Gender

**Converting datatypes**

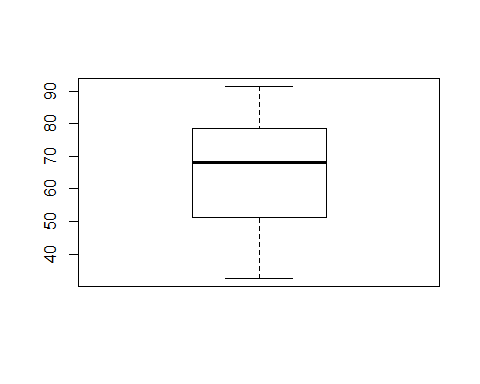
advert$Timestamp <- as.Date(advert$Timestamp, format="Y-%m-%s-%h-%m-%s")  
print(sapply(advert, class))

## Daily.Time.Spent.on.Site Age Area.Income   
## "numeric" "integer" "numeric"   
## Daily.Internet.Usage Ad.Topic.Line City   
## "numeric" "factor" "factor"   
## Gender Country Timestamp   
## "integer" "factor" "Date"   
## Clicked.on.Ad   
## "integer"

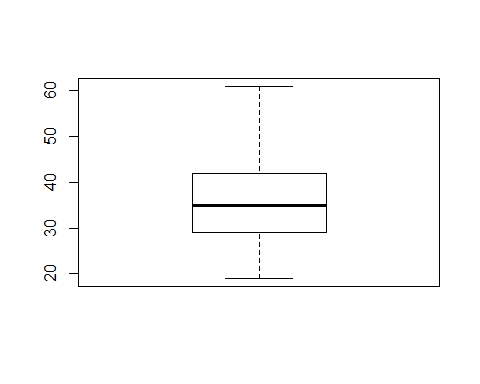
The Timestamp datatype was converted from Factor to Date.

**Check for outliers** This is done on the numeric columns using a boxplot Outliers are shown by plots outside the whisker

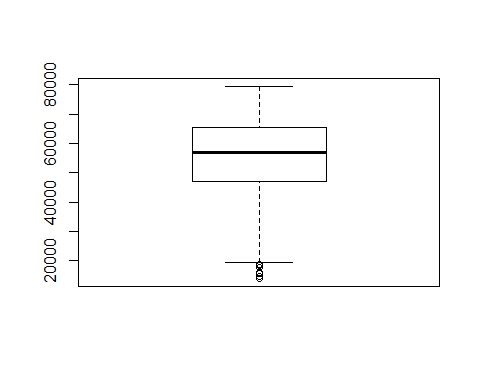
boxplot(advert$Daily.Time.Spent.on.Site, echo=FALSE)

 There are no outliers in the DailyTimeSpentOnSite column

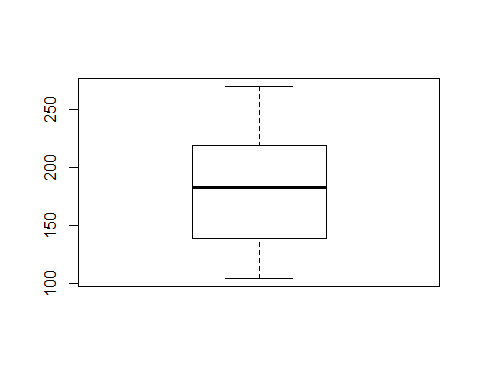
boxplot(advert$Age)

 No outliers o the Age column

boxplot(advert$Area.Income)

 There are someoutliers in the AreaIncome Column but we will not them because that can e a determining factor on the people who click on Ads

boxplot(advert$Daily.Internet.Usage)

 No outliers in the DailyInternetUsage column

## Get Summary statistics of the dataset

summary(advert)

## 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 Gender Country Timestamp   
## Lisamouth : 3 Min. :0.000 Czech Republic: 9 Min. :NA   
## Williamsport : 3 1st Qu.:0.000 France : 9 1st Qu.:NA   
## Benjaminchester: 2 Median :0.000 Afghanistan : 8 Median :NA   
## East John : 2 Mean :0.481 Australia : 8 Mean :NA   
## East Timothy : 2 3rd Qu.:1.000 Cyprus : 8 3rd Qu.:NA   
## Johnstad : 2 Max. :1.000 Greece : 8 Max. :NA   
## (Other) :986 (Other) :950 NA's :1000   
## Clicked.on.Ad  
## Min. :0.0   
## 1st Qu.:0.0   
## Median :0.5   
## Mean :0.5   
## 3rd Qu.:1.0   
## Max. :1.0   
##

Summary statistics gives us 5 key variables needed in analysis.

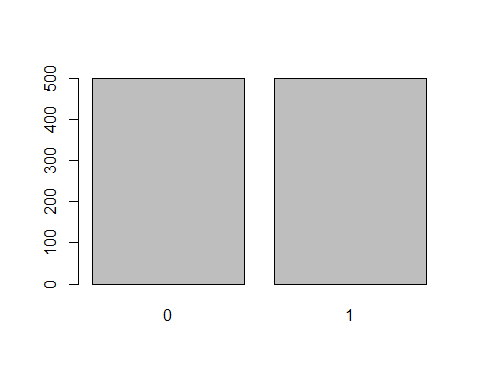
Min: Is the minimum value in the specified column 1st Qu: It splits the lower 25% of data and hence it is the middle value of the lower half Median: Is the midpoint of the data in the specific column Mean: The sum of all the values in the specific column divided by the number of values in the column 3rd Qu: Is the median of the upper half of the data in the particular column Max: It is the maximum value in that column

# Exploratory Data Analysis

## Univariate Analysis

**a)Target variable(ClickedOnAd)**

Clicked.On.Ad\_freq <- table(advert$Clicked.on.Ad)  
barplot(Clicked.On.Ad\_freq)

 0 : Did not click on the Ad 1: Clicked on the Ad We observe that the data is balanced since both are at 500

**b)DailyTimeSpentOnSite** Mean

mean(advert$Daily.Time.Spent.on.Site)

## [1] 65.0002

The average time spent on the site was 65 minutes

Minimum and Maximum timespent on the site

max(advert$Daily.Time.Spent.on.Site)

## [1] 91.43

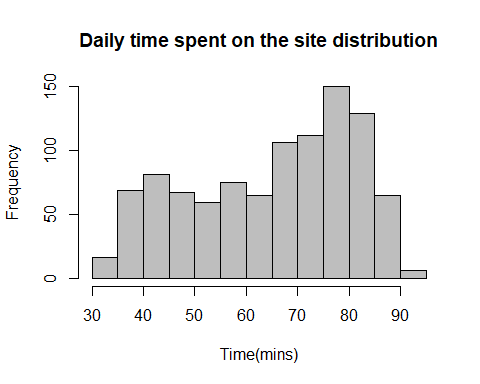
min(advert$Daily.Time.Spent.on.Site)

## [1] 32.6

Max time spent on the site was 91.43 minutes with a Min of 32.6 minutes

**A visualization of the column**

hist(advert$Daily.Time.Spent.on.Site, main = "Daily time spent on the site distribution", xlab = "Time(mins)", col = "gray")

 The data is skewed to the right meaning that people actually spent more time on the site(65-90 mins per day)

**c) Age** Mean

mean(advert$Age)

## [1] 36.009

The people who visited the site where averagely 36 years of age

Minimum and Maximum Age

max(advert$Age)

## [1] 61

min(advert$Age)

## [1] 19

The age bracket was from 19 - 61 which means that the youngest person who visited the site was 19 years with the eldest being 61 years of age

*A Quartile distribution of the Age column*

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

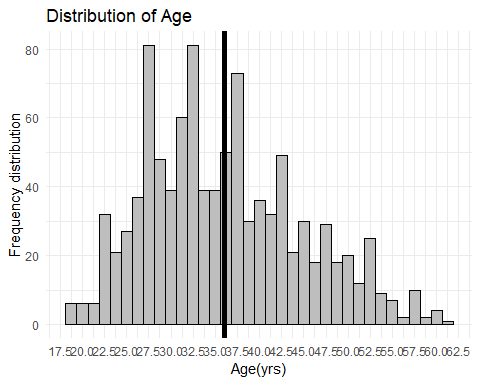
library(pander)  
advert %>%  
 summarize(variable = "Age",  
 q0.25 = quantile(Age, 0.25),  
 q0.50 = quantile(Age, 0.50),  
 q0.75 = quantile(Age, 0.75),  
 q1 = quantile(Age, 1)) %>%  
 pander

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| variable | q0.25 | q0.50 | q0.75 | q1 |
| Age | 29 | 35 | 42 | 61 |

The above distribution of quartiles shows the range of the age column

**A plot of the age column**

library(ggplot2)  
library(dplyr)  
  
advert %>%  
 ggplot(aes(Age)) +  
 geom\_histogram(binwidth = 1.25, color="black",fill="grey") +  
 geom\_vline(xintercept = mean(advert$Age), lwd=2) +  
 labs(title = "Distribution of Age",  
 x = "Age(yrs)",  
 y = "Frequency distribution") +  
 theme\_minimal() +  
 scale\_x\_continuous(breaks = seq(7.5,100,2.5))

 The distribution of the age column is almost normally distributed. The black line intercepting the x-axis shows the position of the mean age value.

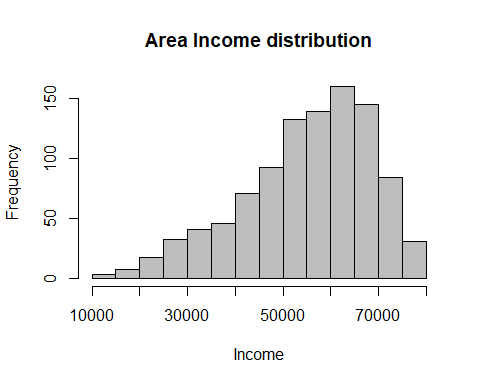
**d)AreaIncome Column** Mean and Standard deviation

library(pander)  
advert %>%  
 summarize(variable = "Area.Income", mean\_income = mean(Area.Income), st\_dev\_Income = sd(Area.Income)) %>%  
 pander

|  |  |  |
| --- | --- | --- |
| variable | mean\_income | st\_dev\_Income |
| Area.Income | 55000 | 13415 |

The average area income was 55,000 with a standard deviation of 13,415

hist(advert$Area.Income, main = "Area Income distribution", xlab = "Income", col = "gray")

 The distribution is skewed to the right meaning that income was high

**e) DailyInternetUsage Column** Mean and Standard deviation

advert %>%  
 summarize(variable = "DailyInternetUsage", Average\_Internet = mean(Daily.Internet.Usage), st\_dev\_Internet = sd(Daily.Internet.Usage)) %>%  
 pander

|  |  |  |
| --- | --- | --- |
| variable | Average\_Internet | st\_dev\_Internet |
| DailyInternetUsage | 180 | 43.9 |

People averagely used 180 mbs perday on that particular site with a standard deviation of 43.9 mbs

Maximum and Minimum DailyInternetUsage

min(advert$Daily.Internet.Usage)

## [1] 104.78

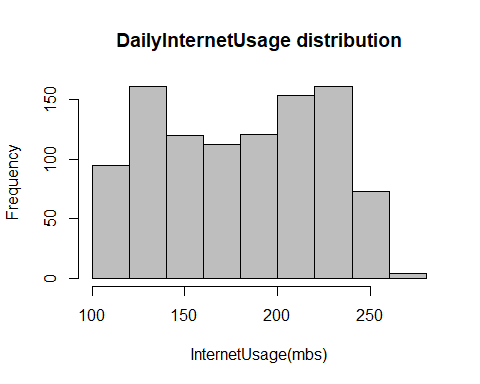
max(advert$Daily.Internet.Usage)

## [1] 269.96

The minimum and maximum daily internet usage was 104.78 mbs and 269.96 mbs respectively.

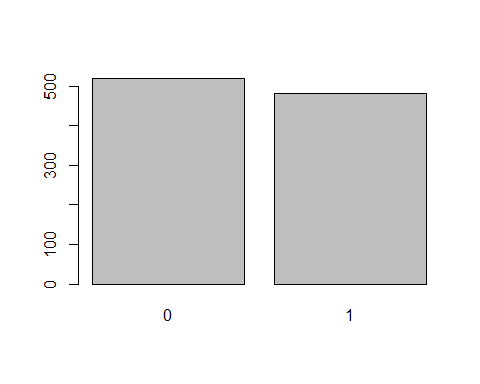
*A visualization of the Daily Internet usage*

hist(advert$Daily.Internet.Usage, main = "DailyInternetUsage distribution", xlab = "InternetUsage(mbs)", col = "gray")



**f) Gender**

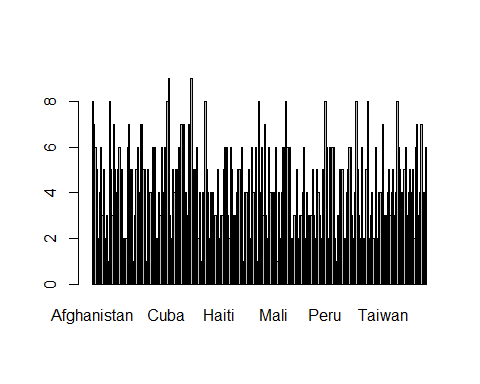
Gender <- advert$Gender  
Gender\_freq <- table(Gender)  
barplot(Gender\_freq)

 0: Female 1: Male

We observe that more females visited the site than males with female being slightly more than 500 and male below 500

**g) Country**

Country <- advert$Country  
country\_freq <- table(Country)  
barplot(country\_freq)

 The distribution above shows 2 countries that projected a frequency of more than 8 which means that those topped the list.

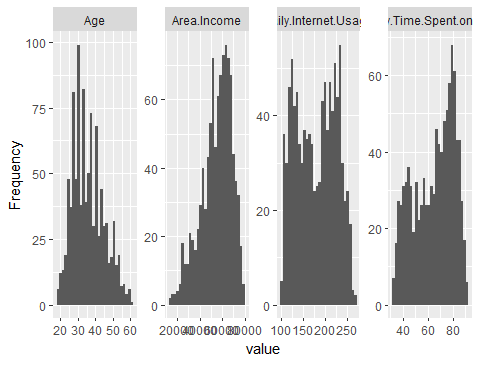
country\_tbl <- advert$Country  
names(table(country\_tbl))[table (country\_tbl)==max(table(country\_tbl))]

## [1] "Czech Republic" "France"

These are the countries that topped the list of the frequent country visits to the website.

**Histograms to enable comparison of the different variables**

library(DataExplorer)  
plot\_histogram(advert)

 ## Bivariate Analysis

*ClickedOnAd Vs DailyTimeSpentOnSite* Covariance Covariance indicates the relationship of two variables whenever one variable changes

ClickedonAd <- advert$Clicked.on.Ad  
DailyTimeSpentonSite <- advert$Daily.Time.Spent.on.Site  
cov(ClickedonAd, DailyTimeSpentonSite )

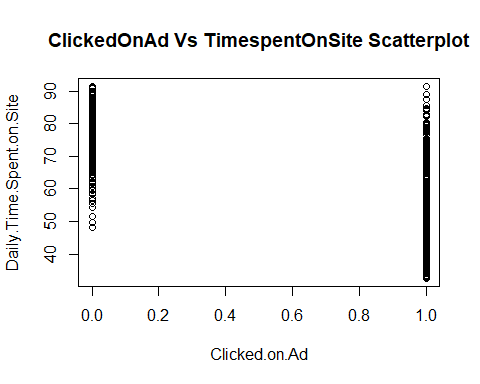
## [1] -5.933143

This means that there’s an inverse relationship

attach(advert)

## The following objects are masked \_by\_ .GlobalEnv:  
##   
## Country, Gender

plot(Clicked.on.Ad, Daily.Time.Spent.on.Site, main="ClickedOnAd Vs TimespentOnSite Scatterplot")

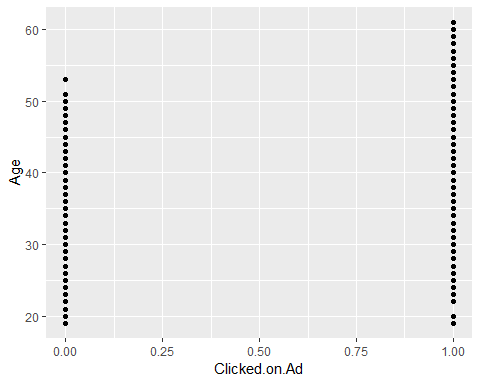
 The scatter plot shows that being on the website longer doesn’t necessarily mean they will click on the Ad.

*Clicked.on.Ad Vs Age*

ClickedonAd <- advert$Clicked.on.Ad  
Age <- advert$Age  
cov(ClickedonAd, Age )

## [1] 2.164665

library(ggplot2)  
ggplot(advert, aes(x = Clicked.on.Ad, y = Age)) +  
 geom\_point()

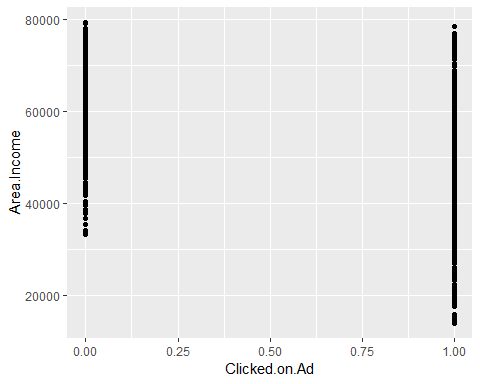
 The plot above shows that there’s a positive relationship because the age plot starts at exactly where the age bracket starts and it goes up directly against the Age column which means that Age determines on whether you will click on an Ad or not.

*ClickedOnAd Vs AreaIncome*

ClickedonAd <- advert$Clicked.on.Ad  
AreaIncome <- advert$Area.Income  
cov(ClickedonAd, AreaIncome)

## [1] -3195.989

library(ggplot2)  
ggplot(advert, aes(x = Clicked.on.Ad, y = Area.Income)) +  
 geom\_point()



min(advert$Area.Income)

## [1] 13996.5

The covariance value(-3195.989) indicates that there is a negative relationship between the target variable(clickedOnAd) and AreaIncome. This is seen in the scatter plot in that having the least income value as 13996.5 and having the graph not start at that value means that more AreaIncome does not mean that they will click on the ad.

*ClickedOnAd Vs Gender*

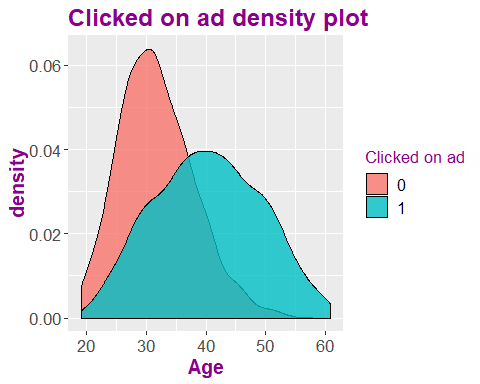
ClickedonAd <- advert$Clicked.on.Ad  
Gender <- advert$Gender  
cov(ClickedonAd, Gender)

## [1] -0.00950951

There is a an iverse relationship between the target variable and Gender. This means that Gender does not determine whether one will click an ad or not.

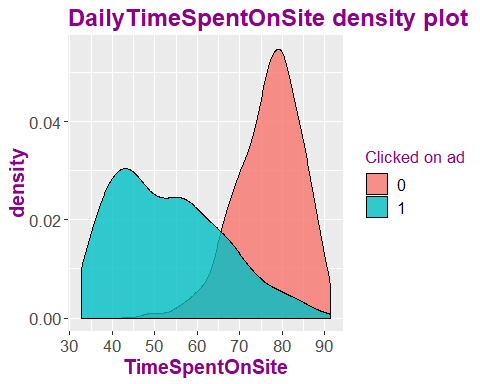
*ClickedOnAd Vs Age density plot* A Density Plot visualises the distribution of data over a continuous interval or time period

library(ggplot2)  
options(repr.plot.width = 13, repr.plot.height = 7)  
R2 = ggplot(data = advert, aes(Age)) +  
 geom\_density(aes(fill=factor(Clicked.on.Ad)), alpha = 0.8) +  
 labs(title = 'Clicked on ad density plot', x = 'Age', fill = 'Clicked on ad') +  
 scale\_color\_brewer(palette = 1) +  
 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 = 0),  
 axis.text.y = element\_text(size = 13),  
 legend.title = element\_text(size = 13, color = 'darkmagenta'),  
 legend.text = element\_text(size = 12))  
plot(R2)



The plot above shows some normal distribution on the people that clicked on the ad meaning it was well ditributed within the age bracket while for the people who did not click on ad having a left skeweness which means that most of the people of the lower age bracket did not click on the ad.

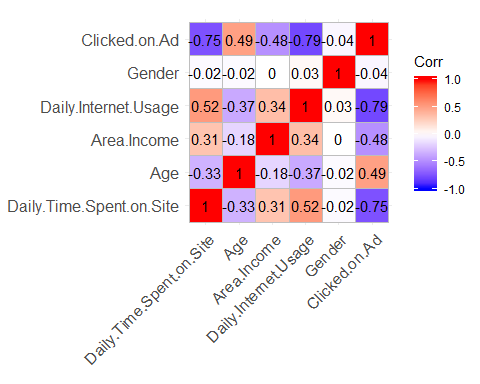
library(ggplot2)  
options(repr.plot.width = 13, repr.plot.height = 7)  
R3 = ggplot(data = advert, aes(Daily.Time.Spent.on.Site)) +  
 geom\_density(aes(fill=factor(Clicked.on.Ad)), alpha = 0.8) +  
 labs(title = 'DailyTimeSpentOnSite density plot', x = 'TimeSpentOnSite', fill = 'Clicked on ad') +  
 scale\_color\_brewer(palette = 1) +  
 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 = 0),  
 axis.text.y = element\_text(size = 13),  
 legend.title = element\_text(size = 13, color = 'darkmagenta'),  
 legend.text = element\_text(size = 12))  
plot(R3)



The people who clicked on the Ad(1) seem to be on the left side of the graph with least time spent on the site with more people people who did not click on the ad(0) being on the right side of the graph where most people spent more time on the site. This means that spending more time on the website does not necessarily mean that someone will click an ad. The people who clicked on the ad graph is more dense between 30-65 mins where density starts to reduce as you get to the 90 min span. This means that most people who clicked on the ad spent 30-65 mins daily.

## Multiple Analysis

library(ggcorrplot)  
library(dplyr)  
advert %>%  
 select\_if(is.numeric) %>%  
 cor %>%   
 ggcorrplot(lab = TRUE)



There is a high correlation between a variable and it’s self. There is a negative correlation between ClickedOnAd and Timespentonsite and Clickedonad and InternetUsage We observe a positive correlation between DailyInternetUsage and Timespentonsite, Daily.Internet.Usage and AreaIncome, AreaIncome and Timespentonsite .

# Modelling

We will use Naive Bayes Classifier > A naive Bayes classifier is an algorithm that uses Bayes’ theorem to classify objects.

#Setting ClickedOnAd variable as categorical  
advert$Clicked.on.Ad <- factor(advert$Clicked.on.Ad, levels = c(0,1), labels = c("False", "True"))

**Splitting the data to train and test data**

library(caret)

## Loading required package: lattice

advert1 <- createDataPartition(y = advert$Clicked.on.Ad,p = 0.75,list = FALSE)  
training <- advert[advert1,]  
testing <- advert[-advert1,]

# Checking dimensions of the split  
  
prop.table(table(advert$Clicked.on.Ad)) \* 100

##   
## False True   
## 50 50

prop.table(table(training$Clicked.on.Ad)) \* 100

##   
## False True   
## 50 50

prop.table(table(testing$Clicked.on.Ad)) \* 100

##   
## False True   
## 50 50

The dataset is well split and hence we can continue with the analysis

# Comparing the outcome of the training and testing phase  
# Creating objects x which holds the predictor variables and y which holds the response variables  
  
x = training[,-c(5,6,8,9,10)]  
y = training$Clicked.on.Ad

# Loading our inbuilt e1071 package that holds the Naive Bayes function.  
library(e1071)

We observe that the model is able to predict correctly a total of 243 observations with only 7 wrong predictions. This is confirmed by the 97.2% model accuracy projected above.