Part 1 crptography course project

Margaret Gathoni

Introduction

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. The columns in the data set include: * Daily_Time_Spent_on_Site

```
* Age

* Area_Income

* Daily_Internet_Usage

* Ad_Topic_Line

* City

* Male

* Country

* Time stamp

* Clicked_on_Ad
```

Problem Statement

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.

Metrics of Success

1. Define the question, the metric for success, the context, experimental design taken and the appropriateness of the available data to answer the given question.
2. Find and deal with outliers, anomalies, and missing data within the data set.
3. Perform uni variate and bivariate analysis.
4. Choose the best supervised learning model to help identify which individuals are most likely to click on the ads in the blog.
5. The model should have an accuracy above 90%.

6. From your insights provide a conclusion and recommendation.

Loading our data set

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.1.2
```

```
advert <- fread('http://bit.ly/IPAdvertisingData')
```

Now let's preview our dataset

First six rows

```
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
```

Will check the class or datatypes in the data set

```
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>
```

Lets check number of rows and columns

```
dim(advert)
```

```
## [1] 1000   10
```

Data set attributes

```
class(advert)
```

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

Data Cleaning and Data Preparation

1. checking for nulls/missing values

```
is.null(advert)
```

```
## [1] FALSE
```

we don't have null values.

2. Checking for duplicates

```
duplicated_rows <- advert[duplicated(advert),]
duplicated_rows
```

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

We have no duplicates

3. Checking for outliers

We have four numeric columns so will check outliers in them

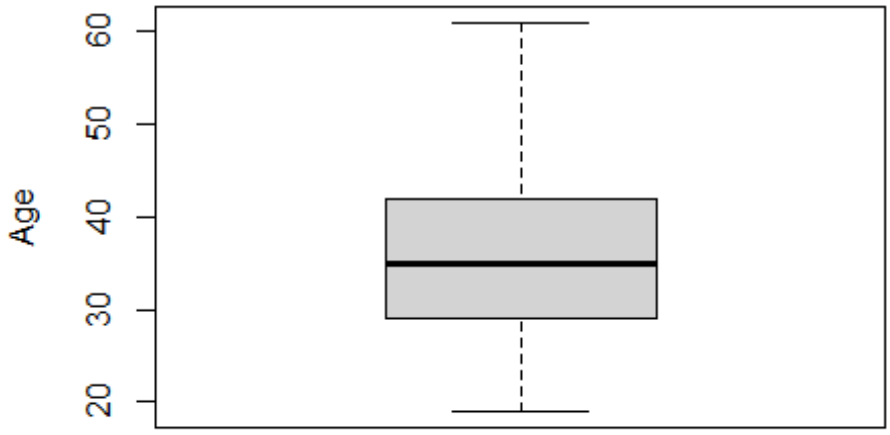
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.3

## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'readr' was built under R version 4.1.2

## Warning: package 'purrr' was built under R version 4.1.2

## Warning: package 'dplyr' was built under R version 4.1.3

## Warning: package 'stringr' was built under R version 4.1.1

## Warning: package 'forcats' was built under R version 4.1.2

## -- 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()
```

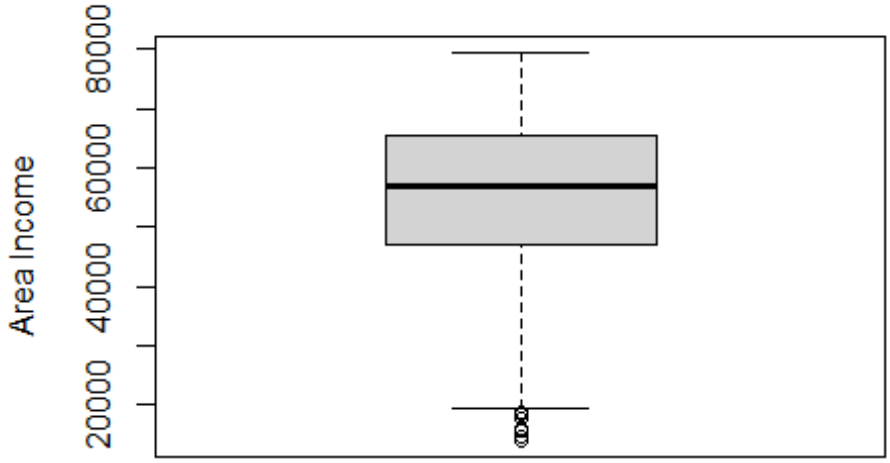
Let's see if we have any outliers in age, as the boxplot shows we have no outliers

```
boxplot(advert$Age, ylab = "Age")
```



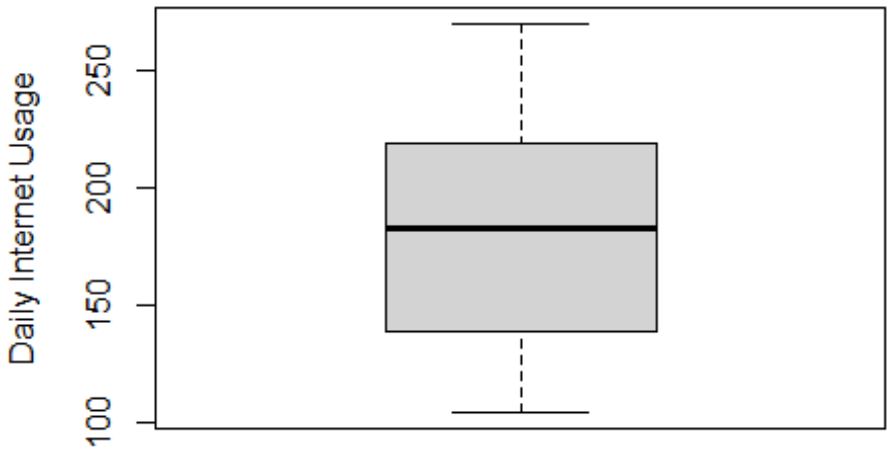
Let's see if we have outliers in area income

```
boxplot(advert$`Area Income`, ylab = "Area Income")
```



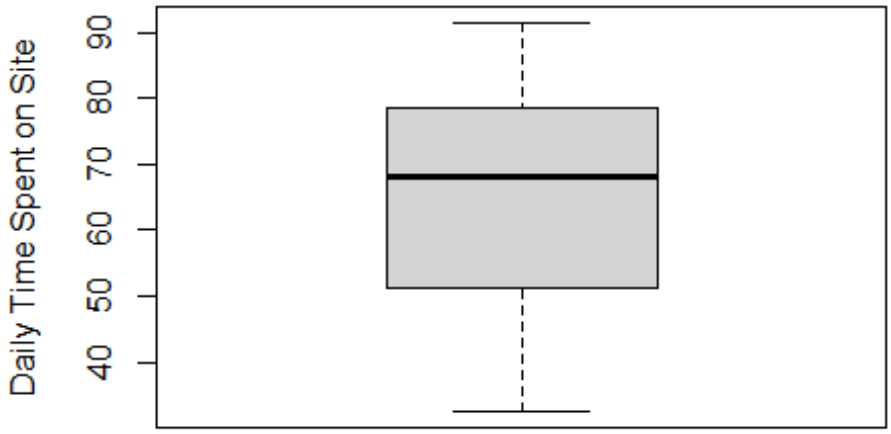
Let's check for outliers in Daily internal usage, we have no outliers as the graph shows

```
boxplot(advert$`Daily Internet Usage`, ylab = "Daily Internet Usage")
```



Let's check for outliers inthe last numeric column

```
boxplot(advert$`Daily Time Spent on Site`, ylab = "Daily Time Spent on Site")
```



Exploratory Data Analysis

1. Uni variate Analysis

Now well move to exploratory data analysis

Measures of central Tendancy

Age

a. Let's see the mean age of the the audience in our client's blog

```
age.mean <- mean(advert$Age)
age.mean
```

```
## [1] 36.009
```

The mean age of most audience is around the age of 36 years

b. Let's see the age range

```
age.range <- range(advert$Age)
age.range
```

```
## [1] 19 61
```

We can see most of the audience age range is between 19 to 61 years.

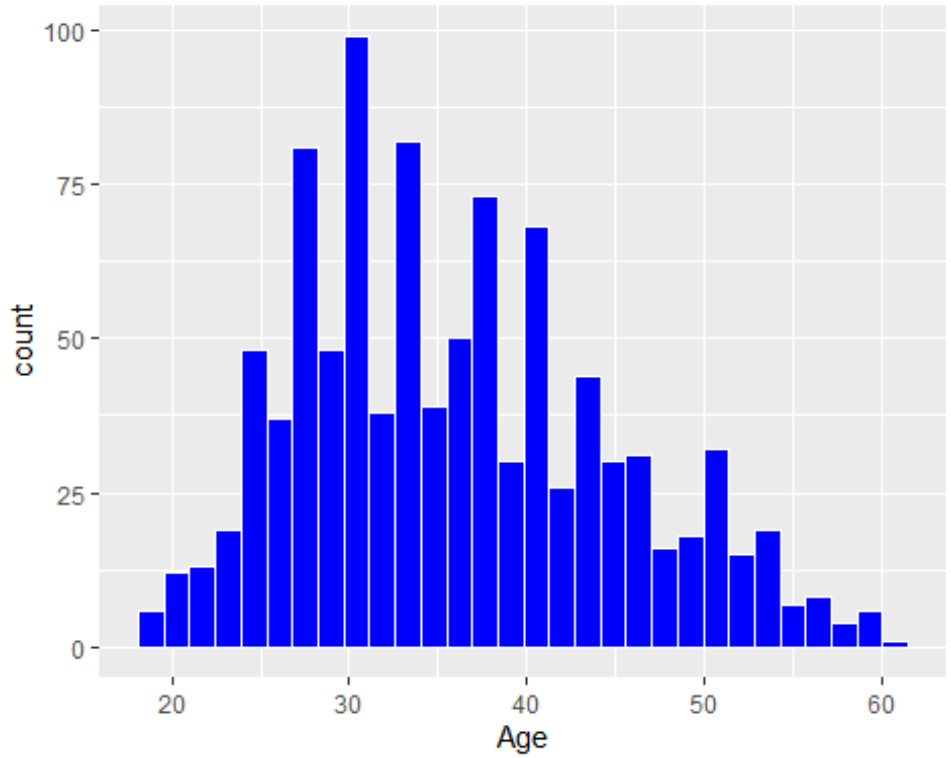
c. Let's see what most of the audiences age bracket is

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

age.mode <- getmode(advert$Age)
age.mode

## [1] 31
```

Most of the audience are 31 years

Visualizing these results

```
ggplot(advert, aes(x=Age)) + geom_histogram(color="white", fill="blue")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



d. Let's check country where most of the audiences come from

```
country.mode <- getmode(advert$Country)
country.mode

## [1] "Czech Republic"
```

Most audiences come from Czech Republic.

e. Let's see country with the minimum hits onthe ads

```
country.min <- min(advert$Country)
country.min
```

```
## [1] "Afghanistan"
```

f. Let's also check the city with most audiences

```
city.mode <- getmode(advert$City)
city.mode
```

```
## [1] "Lisamouth"
```

g. Let's also check the city with least audiences

```
city.min <- min(advert$City)
city.min
```

```
## [1] "Adamsbury"
```

h. We can also check the frequent Ad line topic

```
topic.mode <- getmode(advert$`Ad Topic Line`)
topic.mode
```

```
## [1] "Cloned 5thgeneration orchestration"
```

i. Let's also check for least frequent ad topic line

```
topic.min <- min(advert$`Ad Topic Line`)
topic.min
```

```
## [1] "Adaptive 24hour Graphic Interface"
```

j. Let's see mean daily time spent on site

```
time.mean <- mean(advert$`Daily Time Spent on Site`)
time.mean
```

```
## [1] 65.0002
```

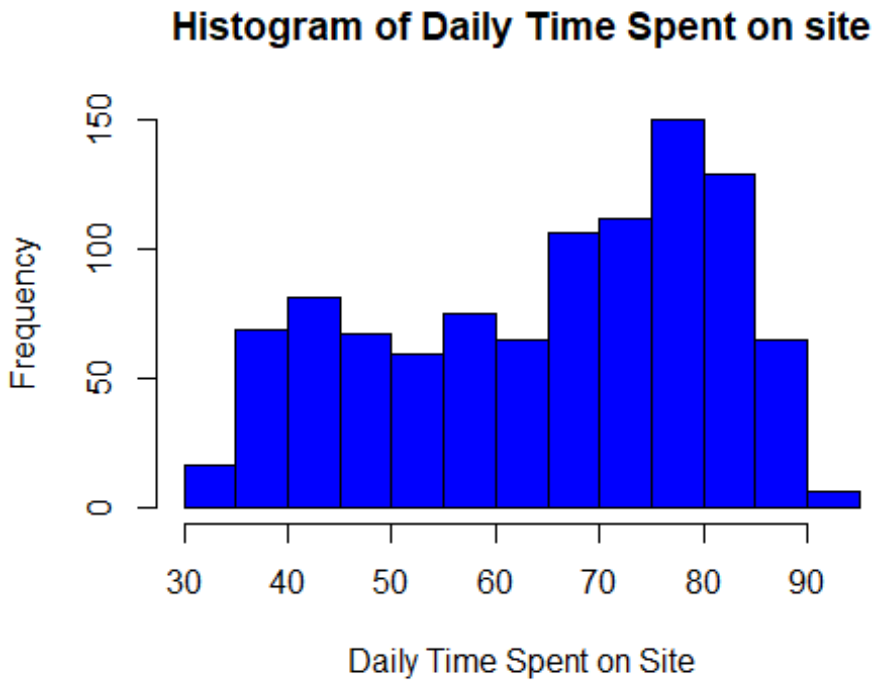
k. Let's check the range of time spent on site daily

```
time.range <- range(advert$`Daily Time Spent on Site`)
time.range
```

```
## [1] 32.60 91.43
```

Will visualize these in a histogram

```
hist((advert$`Daily Time Spent on Site`),
main = "Histogram of Daily Time Spent on site",
     xlab = 'Daily Time Spent on Site',
     ylab = 'Frequency',
     col = "blue")
```



l. determining mean area income

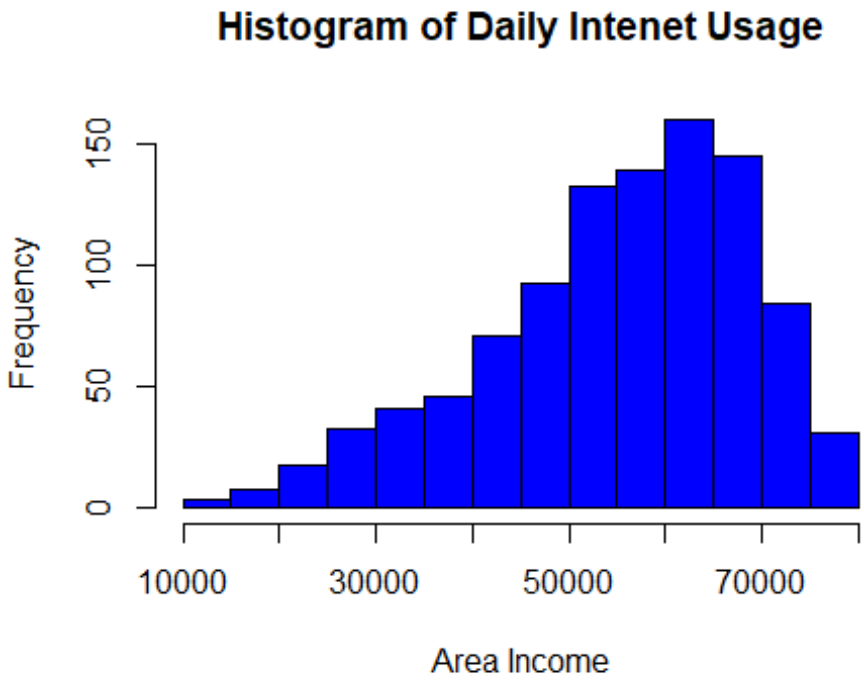
```
area.mean <- mean(advert$`Area Income`)
area.mean

## [1] 55000
```

m. Range of area income

```
area.range <- range(advert$`Area Income`)
area.range

## [1] 13996.5 79484.8
```

Visualize these results

```
hist((advert$`Area Income`),
main = "Histogram of Daily Intenet Usage",
     xlab = 'Area Income',
     ylab = 'Frequency',
     col = "blue")
```



n. Determining mean Daily internet usage

```
usage.mean <- mean(advert$`Daily Internet Usage`)
usage.mean
```

```
## [1] 180.0001
```

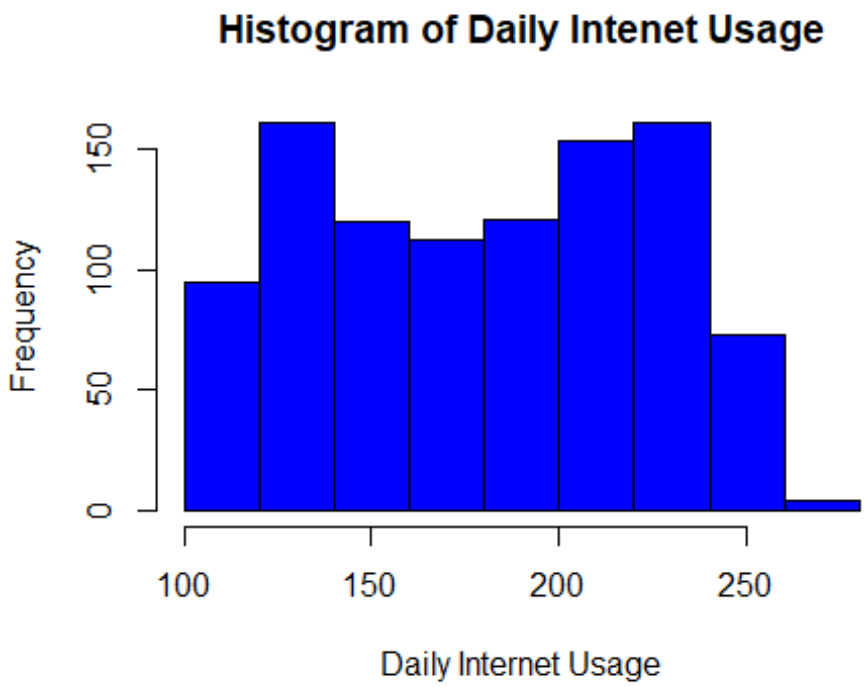
o. range of internet usage daily

```
usage.range <- range(advert$`Daily Internet Usage`)
usage.range
```

```
## [1] 104.78 269.96
```

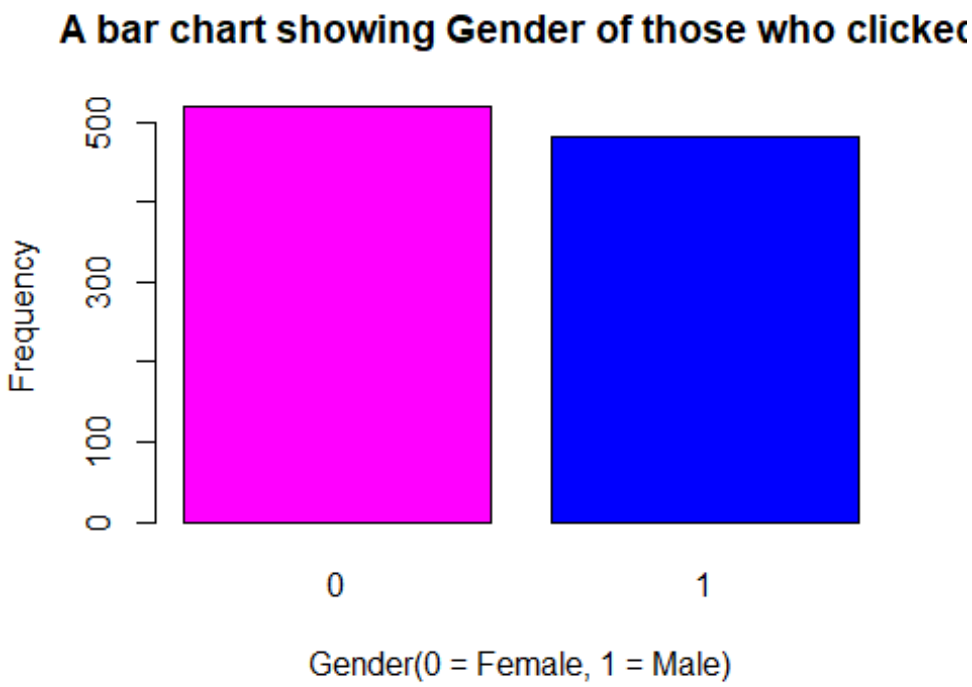
Will visualize this below

```
hist((advert$`Daily Internet Usage`),
main = "Histogram of Daily Intenet Usage",
     xlab = 'Daily Internet Usage',
     ylab = 'Frequency',
     col = "blue")
```



p. Now let's visualize gender distribution

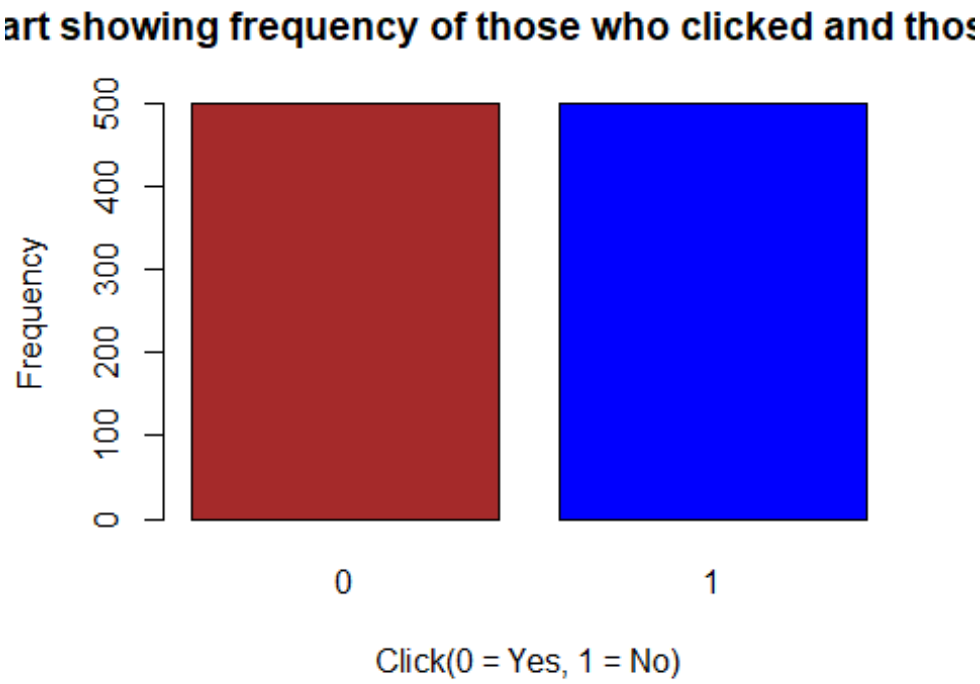
```
gender <- (advert$Male)
gender.frequency <- table(gender)
gender.frequency

## gender
##   0   1
## 519 481

barplot(gender.frequency,
  main="A bar chart showing Gender of those who clicked",
  xlab="Gender(0 = Female, 1 = Male)",
  ylab = "Frequency",
  col=c("magenta","blue"),
  )
```



We see there are more females than male

q. Now let's visualize clicks distribution

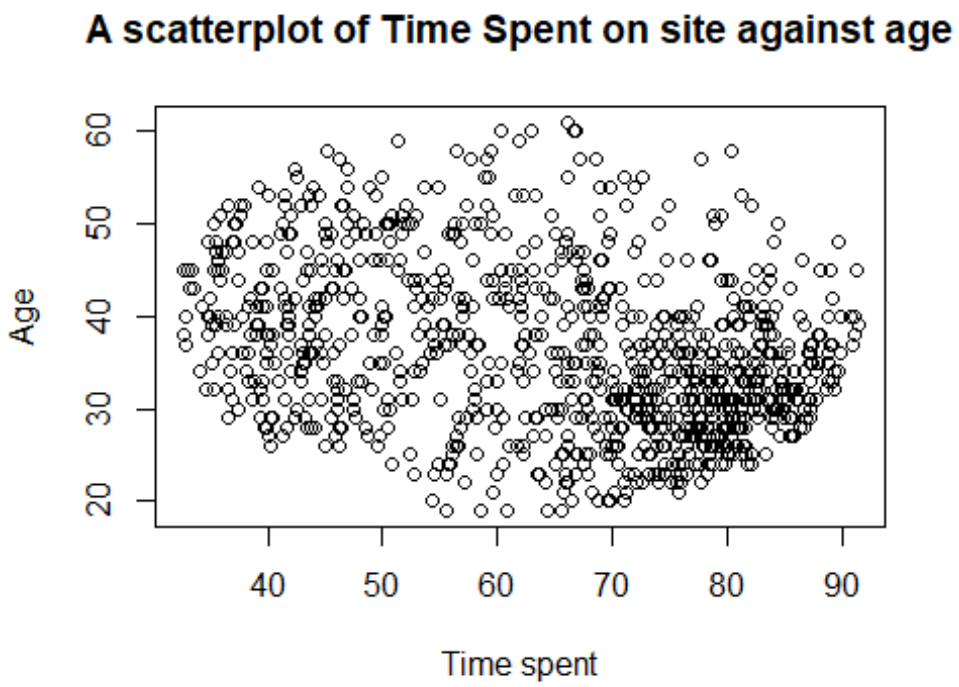
```
click <- (advert$`Clicked on Ad`)
click.frequency <- table(click)
click.frequency

## click
##   0   1
## 500 500

barplot(click.frequency,
  main="A bar chart showing frequency of those who clicked and those who
didn't",
  xlab="Click(0 = Yes, 1 = No)",
  ylab = "Frequency",
  col=c("brown","blue"),
  )
```



Bivariate and Multivariate Analysis

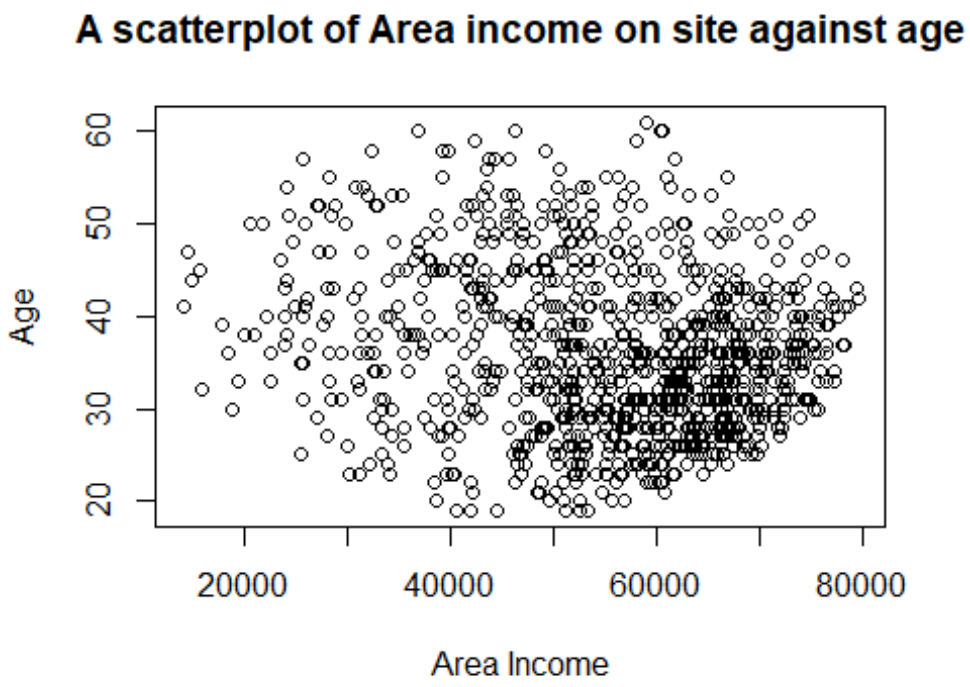
Here will be comparing two or more variables to try and understand their comparison

```
plot((advert$`Daily Time Spent on Site`), (advert$Age),
     main = "A scatterplot of Time Spent on site against age",
     xlab = 'Time spent',
     ylab = 'Age')
```



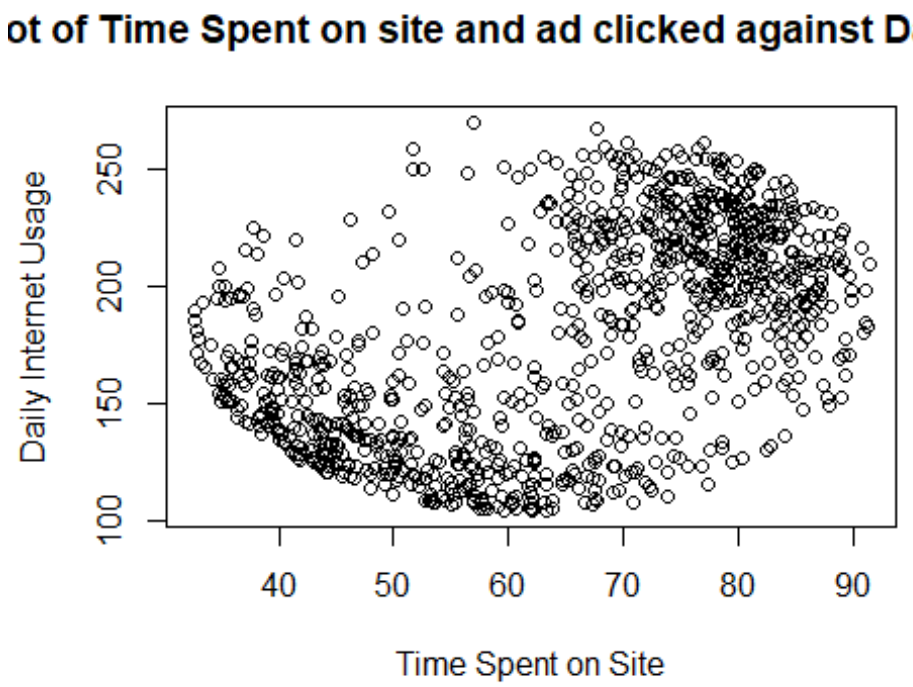
We see there is concentration of Time spent daily on the site in relation to Age is concentrated on around 30s age bracket

```
plot((advert$`Area Income`), (advert$Age),
     main = "A scatterplot of Area income on site against age",
     xlab = 'Area Income',
     ylab = 'Age')
```



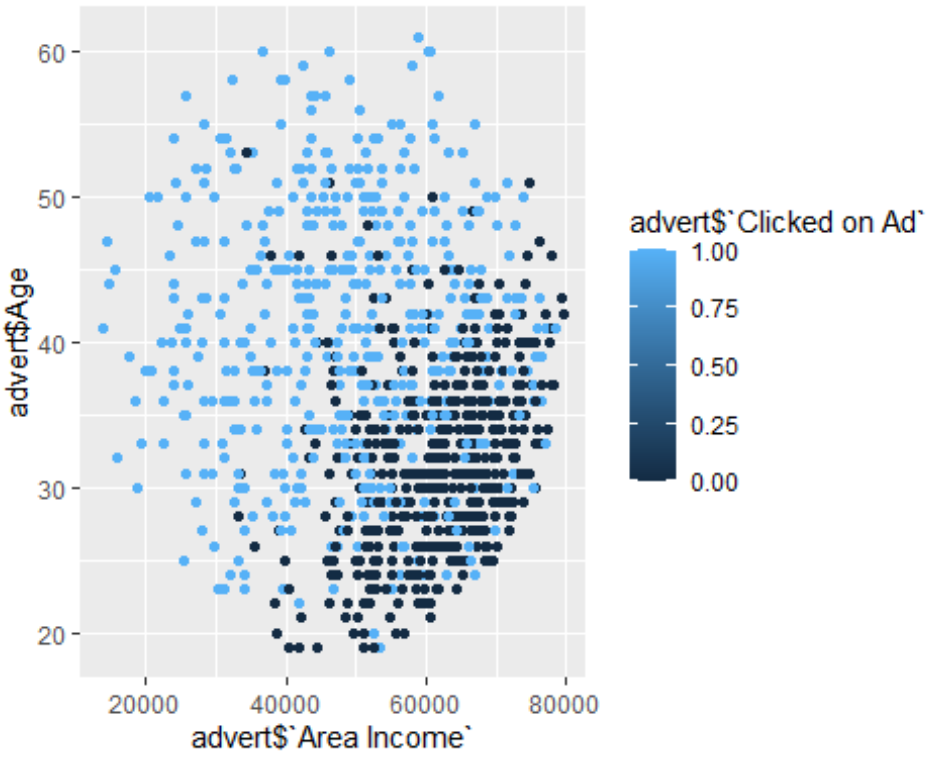
Most people concentrate around 50000 to 70000 area income for the majority age bracket

```
plot((advert$`Daily Time Spent on Site`), (advert$`Daily Internet Usage`),
     main = "A scatterplot of Time Spent on site and ad clicked against Daily
Internet Usage",
     xlab = 'Time Spent on Site',
     ylab = 'Daily Internet Usage')
```



Internet usage and time spent has a linear correlation

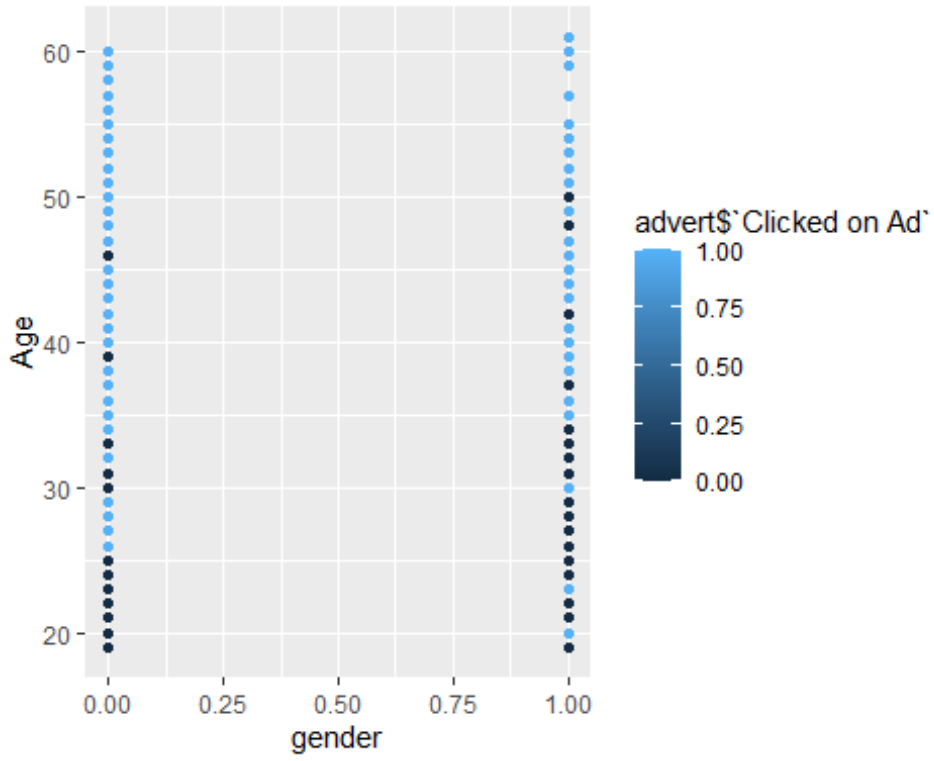
```
ggplot(advert,aes(x=advert$`Area Income`,y=advert$Age,col=advert$`Clicked on Ad`))+geom_point(aes(color=advert$`Clicked on Ad`))
```



The graph shows us that The higher the income the more the clicks but also mostly concentrated around the 30s age bracket.

Let's see click vs gender

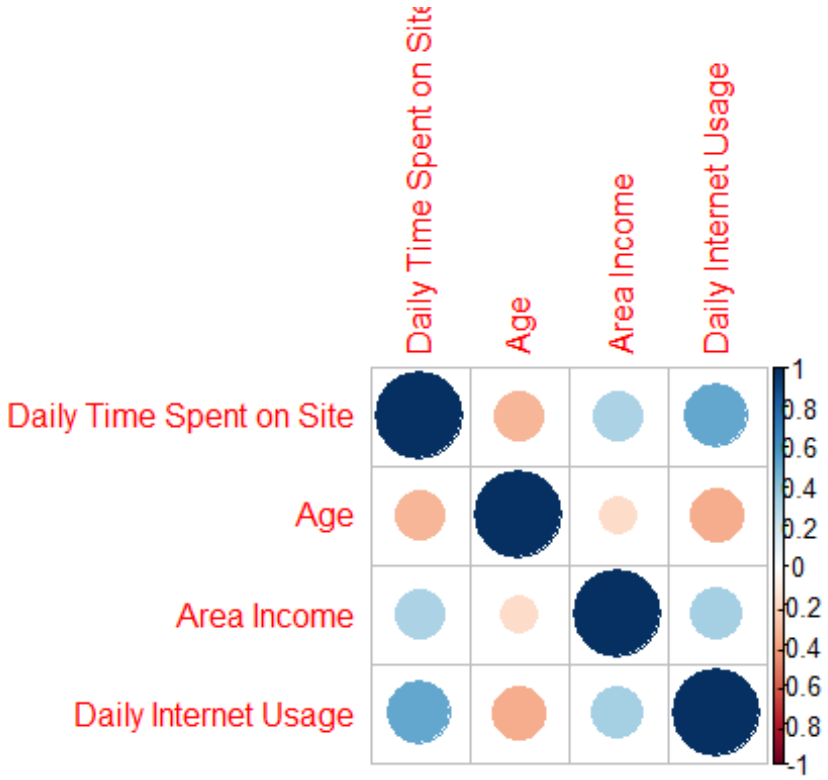
```
ggplot(advert,aes(x=gender,y=Age,col=advert$`Clicked on
Ad`))+geom_point(aes(color=advert$`Clicked on Ad`))
```



Despite most audience visiting the site being female , most of the those who click the add are male

Let.s check for correlation

```
library(corrplot)

## Warning: package 'corrplot' was built under R version 4.1.3

## corrplot 0.92 loaded

numeric <- advert %>%
  select_if(is.numeric) %>%
  select("Daily Time Spent on Site", "Age", "Area Income", "Daily Internet
Usage")

corrplot(cor(numeric))
```



There a positive correlation observed inth plot between Daily internet usage vs time spent on internet daily.

There is also some correlation between area income and internet usage and time

Modeling

Creating a supervised learning model to help identify which individuals are most likely to click on the ads in the blog.

Reminding ourselves how the data looks like

```
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
```

A KNN Model

```
library(e1071)

## Warning: package 'e1071' was built under R version 4.1.3

library(caTools)

## Warning: package 'caTools' was built under R version 4.1.3

library(class)

## Warning: package 'class' was built under R version 4.1.3
```

Will drop some column that will not me necessary moving forward

```
advert$`Ad Topic Line` <- NULL
advert$City <- NULL
advert$Country <- NULL
advert$Timestamp <- NULL
head(advert)

##    Daily Time Spent on Site Age Area Income Daily Internet Usage Male
## 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
## 6:                    59.99  23    59761.56               226.74    1
##    Clicked on Ad
## 1:             0
## 2:             0
## 3:             0
## 4:             0
## 5:             0
## 6:             0
```

A. Splitting data into train and test

```
split <- sample.split(advert, SplitRatio = 0.8)
train <- subset.data.frame(advert, split == "TRUE")
test <- subset.data.frame(advert, split == "FALSE")
```

b. Checking the records for train and test

```
dim(train)
```

```
## [1] 667   6
```

```
dim(test)
```

```
## [1] 333   6
```

c. Feature Scaling

```
train_scale <- scale(train[, 1:5])
test_scale <- scale(test[, 1:5])
```

d. Fitting our KNN Model to the training data set

```
classifier_KNN <- knn(train = train_scale,
                      test = test_scale,
                      cl = train$`Clicked on Ad`,
                      k = 5)

classifier_KNN
```

```
##   [1] 0 0 0 0 1 0 0 0 1 0 1 0 1 0 1 0 0 1 1 1 0 0 1 0 0 0 0 1 1 1 1 0 0 0
0 1 0
##  [38] 1 1 0 1 0 0 1 1 1 1 0 1 1 0 0 0 0 1 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 1
1 1 0
##  [75] 1 1 1 0 0 0 0 1 1 1 0 1 1 1 0 0 0 1 0 1 0 0 0 0 0 0 1 0 1 0 1 0 0 0
0 1 0
## [112] 1 0 1 0 1 0 0 1 0 1 1 0 1 0 1 1 0 1 1 0 0 0 0 0 1 0 1 0 1 1 1 0 0 1
0 0 1
## [149] 0 0 0 0 0 1 1 1 0 0 1 1 1 1 0 1 0 1 0 1 1 0 1 1 1 1 1 0 1 0 0 0 1 1
0 0 0
## [186] 0 1 1 0 0 0 1 0 1 0 1 1 0 0 1 0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 1 0 0 0
0 1 1
## [223] 1 1 0 1 1 1 0 0 0 1 0 0 1 0 0 0 1 0 1 1 0 0 0 0 1 0 1 1 1 0 0 0 0 1 1
0 1 0
## [260] 0 0 1 1 1 1 0 0 1 1 1 1 0 0 0 0 0 1 1 1 1 1 0 1 1 1 1 0 0 0 0 0 0 1 0
0 1 1
## [297] 0 0 1 1 1 0 0 0 1 1 0 1 1 0 1 1 0 0 1 0 1 1 1 0 0 0 1 0 1 1 1 0 0 1 0
0 0 1
## Levels: 0 1
```

e. Confusion Matrix

```
cm <- table(test$`Clicked on Ad`, classifier_KNN)
cm
```

```
##    classifier_KNN
##       0   1
##   0 168   6
##   1  11 148
```

The confusion Matrix above shows that the model was able to correctly identify 172 for class 0 and made 4 wrong prediction for the same class.

f. Evaluating our model

```
misClassError <- mean(classifier_KNN != test$`Clicked on Ad`)
print(paste('Accuracy =', 1-misClassError))

## [1] "Accuracy = 0.948948948948949"
```

Our model has a 95% accuracy.

Conclusion

- The mean age of most audience is 36 years with most of the audience being around age 31 and the range of audience visiting the site is between 19 and 61 years.
- The mean time spent on site is 65 minutes with a range between 32 to 91 minutes on the site.
- The mean Daily Internet Usage is 180 with a range 104 to 269.
- The mean area income is 55000 with a range of 13996 - 79484.
- The country with most audience is Czech Republic and the least was Afghanistan.
- There are more females visiting the site compared to males, however, for those actually clicking the ads the most are males.
- There is a strong correlation between time spent and internet usage on the site which comes out as expected.
- The age of most audiences clicking on the site has a correlation also with the area income i.e. most of those clicking the add around 30 years bracket also has an area income above 50000.
- With our model of choice we got an accuracy of 95 % ## **Recommendation**
- To first answer our stakeholder question of which individuals are most likely to click on her ads: These individuals are male around the age 30 to 35 with an area income above 50000.
- More advertisement need to be done locally however not meaning they should focus only on local audience.
- Most of the those who click on the ad have an area income above 50000, so maybe reevaluate the prices or other ways to attract those in low income areas.