R-Programming_EDA_Week1_IP

RuthNguli

2022-03-18

Defining the Question

To identify which individuals are most likely to click on an online cryptography course ads.

Metric for success

The metric of success will be attained on identifying individuals who click on the ads.

Understanding the business context

Cryptography is an indispensable tool for protecting information in computer systems. A cryptography course teaches how the cryptographic system a works and its real world application.

Experimental Design

Define the question, the metric for success, the context, experimental design taken.

Read and explore the given dataset.

Cleaning Data

Perform Exploratory Data Cleaning (Univariate & Bivariate)

Modelling

Conclusion

Recommendations

1. Reading data

```
# read data from url: http://bit.ly/IPAdvertisingData
# load data
#
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ads <- read.csv("http://bit.ly/IPAdvertisingData")

# preview head of the data
#
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
```

```
# previewing the tail of the data
#
tail(ads)
```

```
##      Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage
## 995                     43.70  28    63126.96               173.01
## 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
## 995         Front-line bifurcated ability  Nicholasland    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
## 995                  Mayotte 2016-04-04 03:57:48             1
## 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 column names
#
colnames(ads)
```

```
##  [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"
```

```
# Checking the data has appropriate data types
#
str(ads)
```

```
## '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 standardi
##  $ 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               : chr  "2016-03-27 00:53:11" "2016-04-04 01:39:02" "2016-03-13 20:35:42"
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

The advertising data has 1000 rows and 10 columns Column Male and Clicked.on.Ad are represented as integer however should be converter to factor as they are represnting categorical variables

Data Cleaning

```
# change data types of Male and Clicked.on.Ad columns from int to factor
#
ads$Male <- as.factor(ads$Male)
ads$Clicked.on.Ad <- as.factor(ads$Clicked.on.Ad)

# confirming if the changes have made successfully
#
str(ads)
```

```
## '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 standardi
##  $ 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               : chr  "2016-03-27 00:53:11" "2016-04-04 01:39:02" "2016-03-13 20:35:42"
##  $ Clicked.on.Ad           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
```

```
#Extracting and creating Date column from timestamp
#
ads$Date <- as.Date(ads$Timestamp)
glimpse(ads)
```

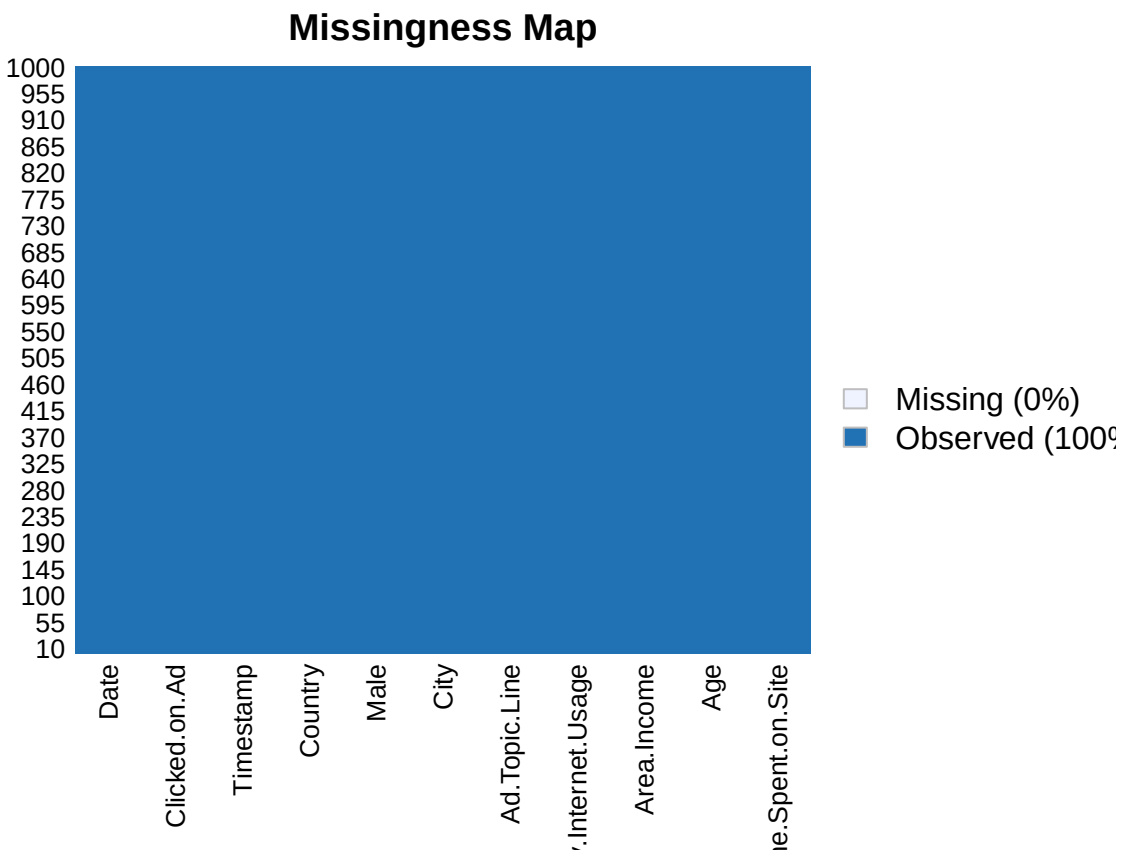
```
## Rows: 1,000
## Columns: 11
## $ Daily.Time.Spent.on.Site <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, 88.~
## $ Age                      <int> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49, 3~
## $ Area.Income              <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 73889~
## $ Daily.Internet.Usage     <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 226.7~
## $ Ad.Topic.Line            <chr> "Cloned 5thgeneration orchestration", "Monito~
## $ City                     <chr> "Wrightburgh", "West Jodi", "Davidton", "West~
## $ Male                     <fct> 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, ~
## $ Country                  <chr> "Tunisia", "Nauru", "San Marino", "Italy", "I~
## $ Timestamp                <chr> "2016-03-27 00:53:11", "2016-04-04 01:39:02",~
## $ Clicked.on.Ad            <fct> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, ~
## $ Date                     <date> 2016-03-27, 2016-04-04, 2016-03-13, 2016-01-~
```

```
# Checking missing values
#
library(Amelia)
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.0, built: 2021-05-26)
## ## Copyright (C) 2005-2022 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
missmap(ads)
```



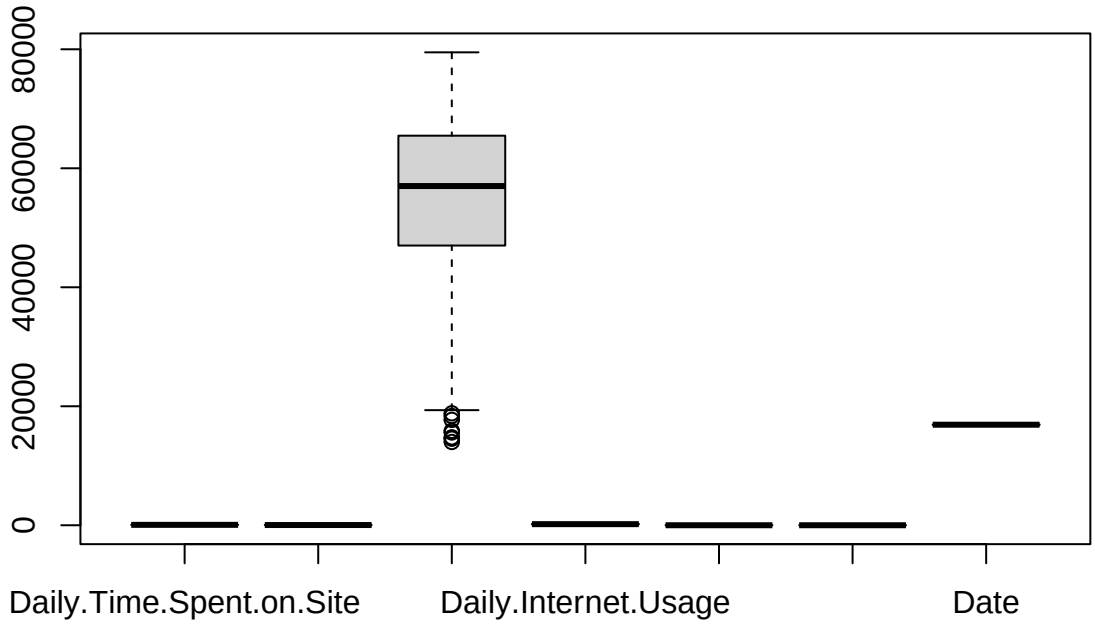
Our dataset does not have any missing values

```
# Checking for any duplicates
#
sum(duplicated(ads))
```

```
## [1] 0
```

There are no duplicates

```
# Checking for outliers
#
non_char <- ads %>% select(Daily.Time.Spent.on.Site, Age, Area.Income, Daily.Internet.Usage, Male,Click
boxplot(non_char)
```



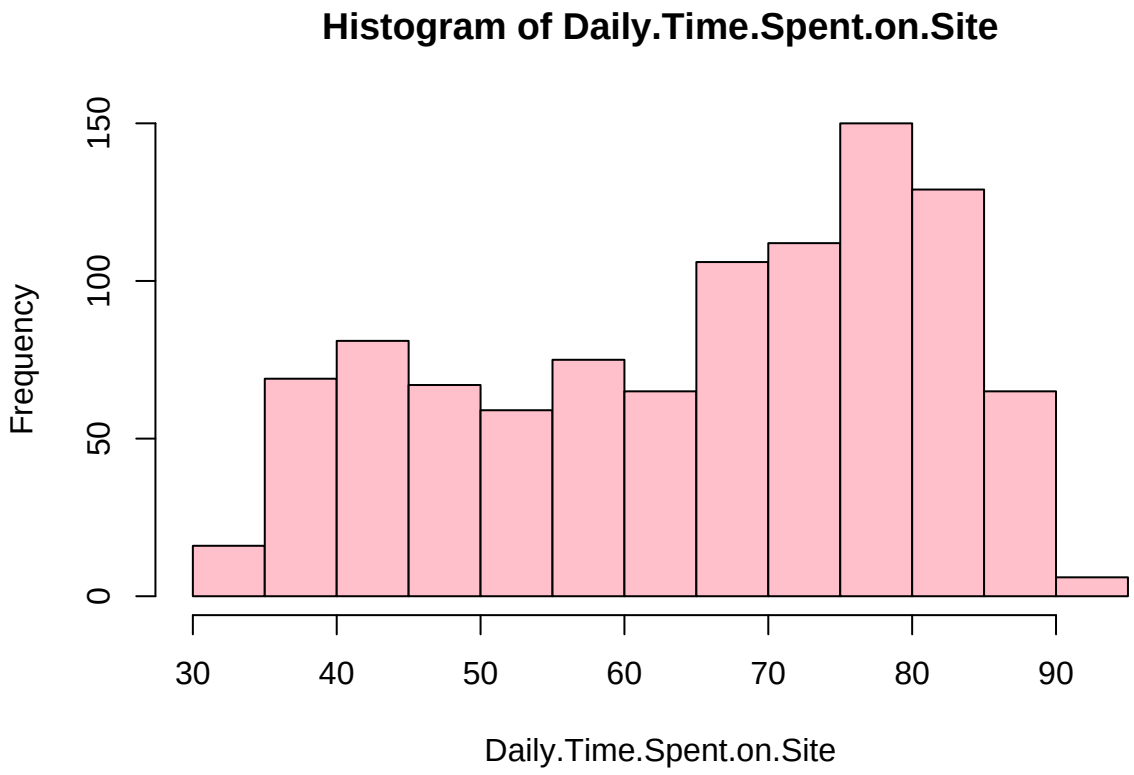
There a few identifiable outliers in Area.Income column, we leave them as they represent real data.

```
# Checking the summarry of dataset
#
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      Date
##  Length:1000        0:500         Min.   :2016-01-01
##  Class :character   1:500         1st Qu.:2016-02-17
##  Mode  :character                 Median :2016-04-07
##                                   Mean   :2016-04-09
##                                   3rd Qu.:2016-05-31
##                                   Max.   :2016-07-24
```

Bivariate analysis

```
# Attaching ads data to R
#
attach(ads)
# Plotting a histogram of Daily Time spent on site
hist(Daily.Time.Spent.on.Site, col='pink')
```

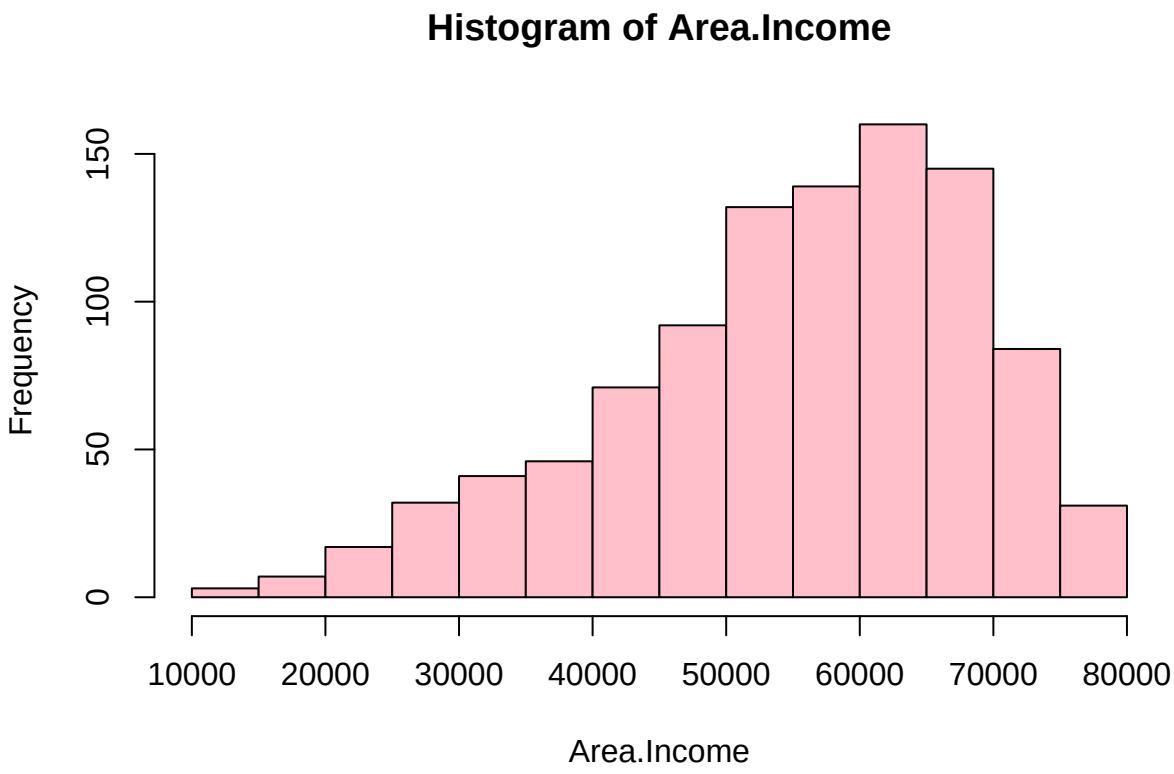


Most time spent on site is between 65 and 80

```
#Plotting a histogram of Age
#
hist(Age, col='pink')
```

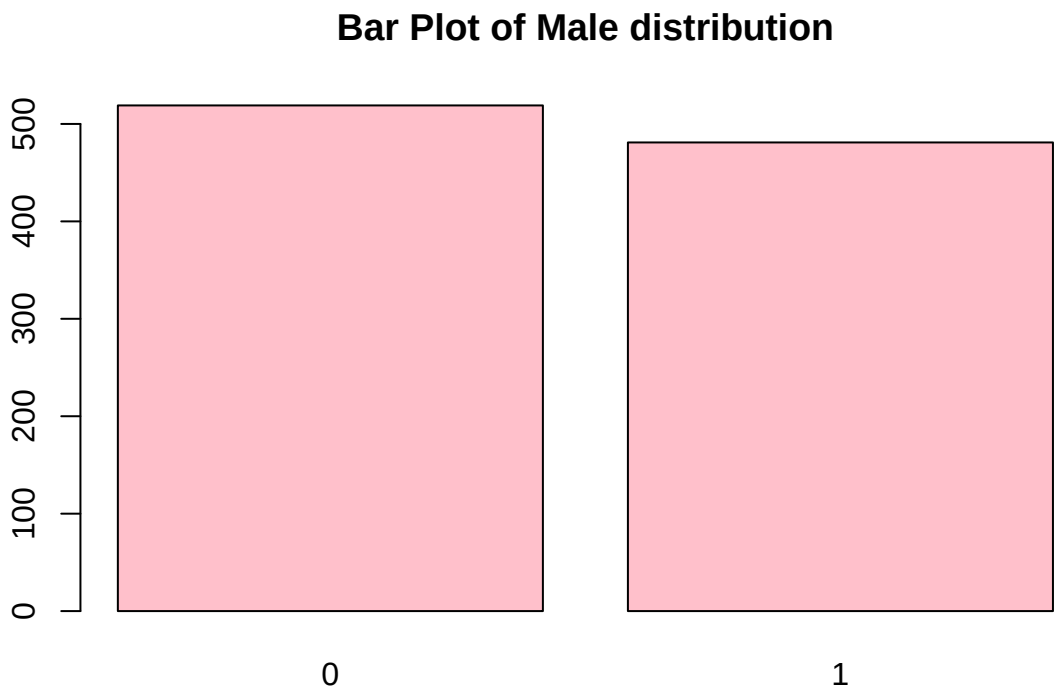
Most participants are aged between 25 and 40 years old

```
# plotting a histogramof Area income
#
hist(Area.Income, col="pink")
```



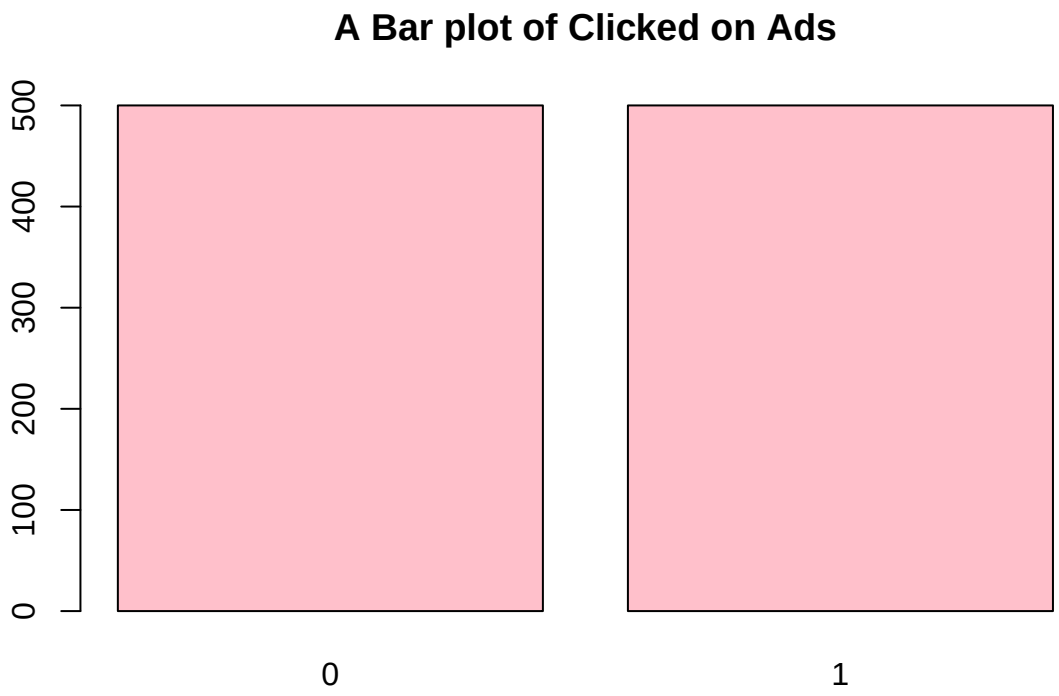
Income is skewed to the left. Most participants had income between 50,000 and 70,000

```
# A bar plot of Male Partipation
#
barplot(table(Male), col="pink", main="Bar Plot of Male distribution")
```



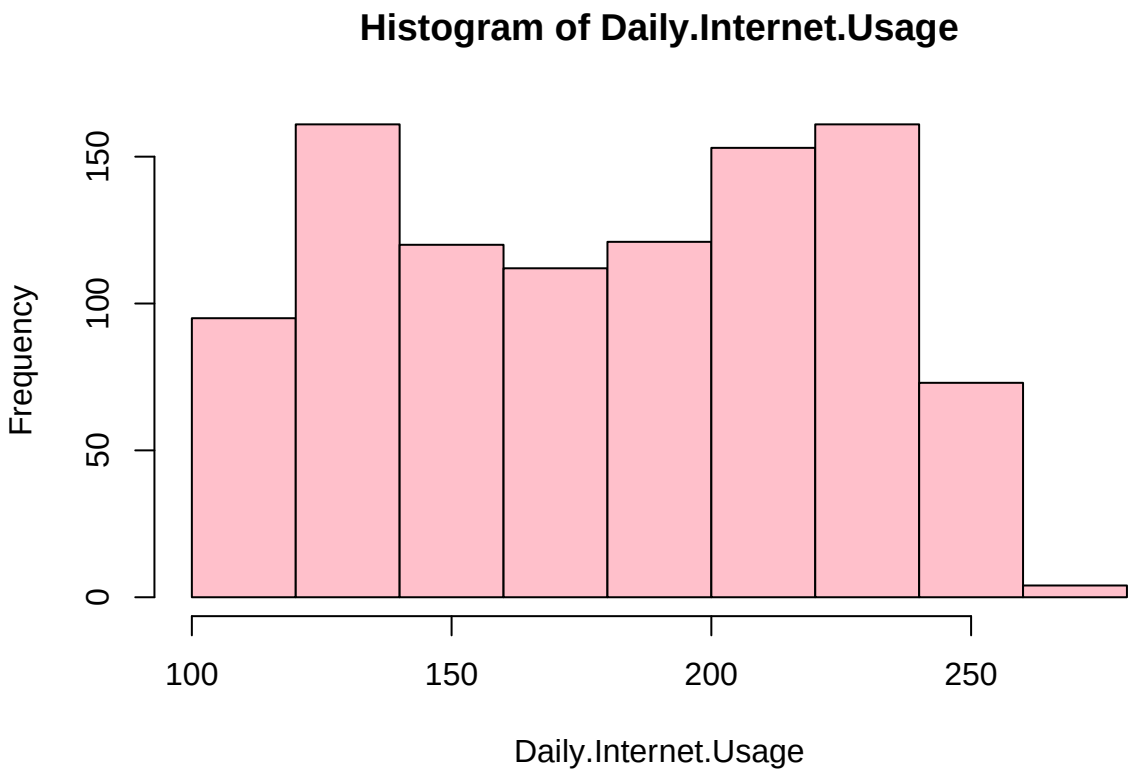
Male participants were slightly fewer than those not male

```
# A bar plot of Clicked on Ads
#
barplot(table( Clicked.on.Ad), col="pink", main="A Bar plot of Clicked on Ads")
```



There is equal distribution between those who clicked and those didn't click on AD

```
# Plotting a Histogram of Daily internet usage
#
hist(Daily.Internet.Usage, col="pink")
```

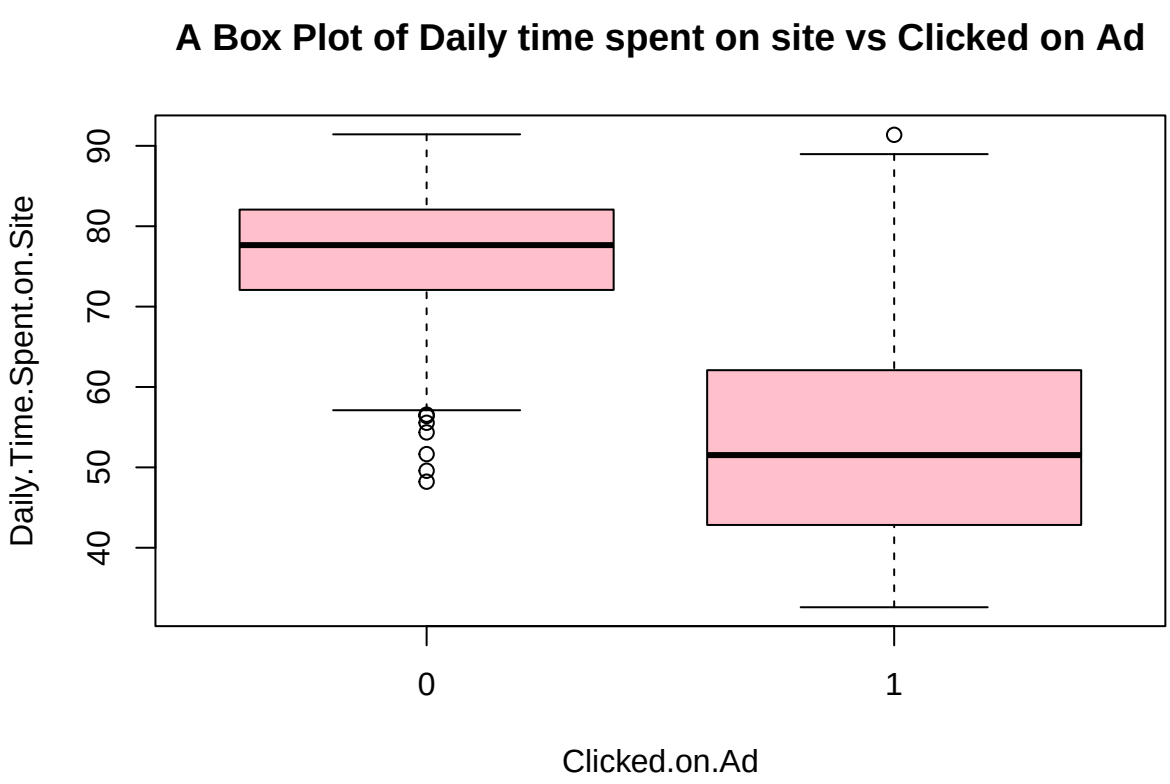


Bivariate Analysis

```
colnames(ads)
```

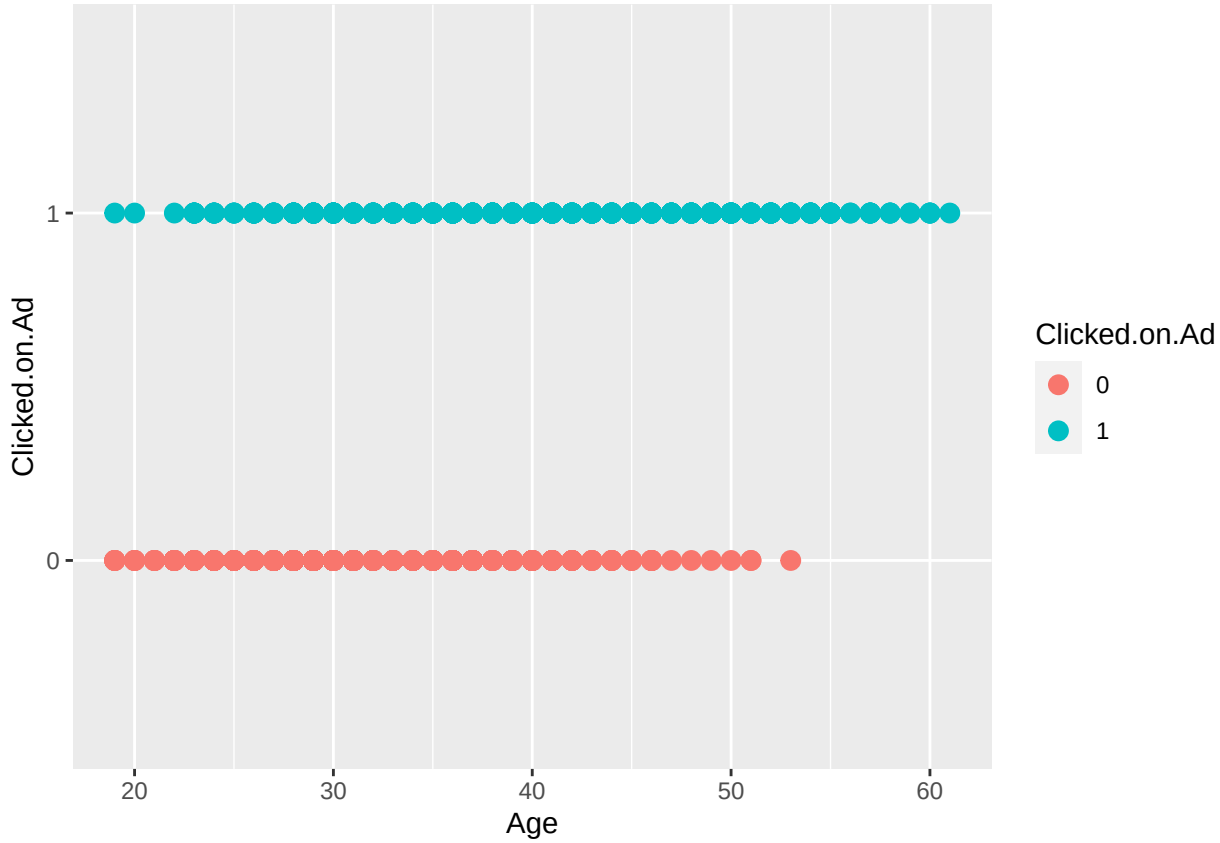
```
##  [1] "Daily.Time.Spent.on.Site" "Age"
##  [3] "Area.Income"              "Daily.Internet.Usage"
##  [5] "Ad.Topic.Line"            "City"
##  [7] "Male"                     "Country"
##  [9] "Timestamp"                "Clicked.on.Ad"
## [11] "Date"
```

```
# A Boxplot of clicked on ad vs Time spent on ad
#
plot(Daily.Time.Spent.on.Site ~ Clicked.on.Ad, data = ads, col="pink", main="A Box Plot of Daily time sp
```



* Most people who clicked on ads did not spent much time on site*

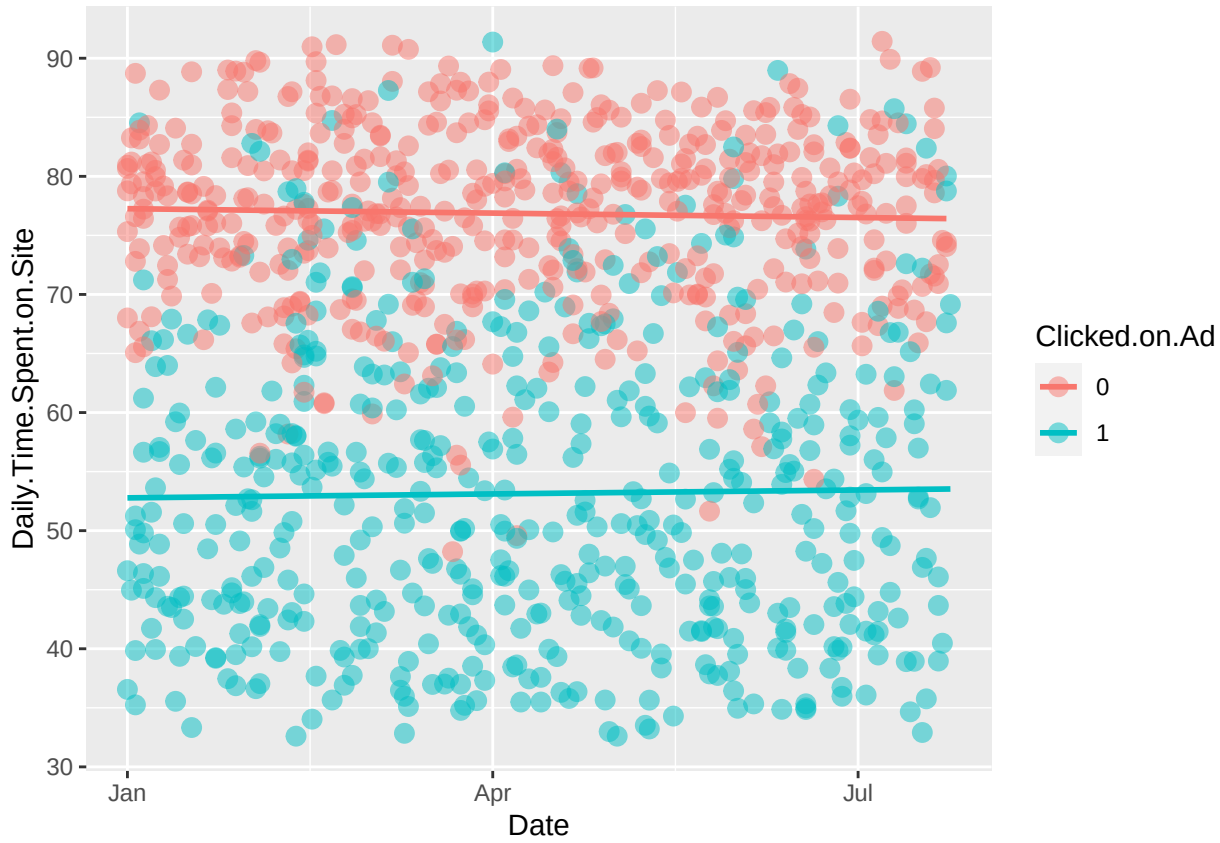
```
# A scatter-plot of Age Vs Clicked on ads
#
ggplot(ads,aes(Age,Clicked.on.Ad, colour= Clicked.on.Ad))+
  geom_point(size=3)
```



Those who clicked on ads are aged between 20 and 60.

```
# A Scatter plot of Date Vs Daily time spent on site
#
ads %>% ggplot(aes(Date,Daily.Time.Spent.on.Site, colour = Clicked.on.Ad))+
  geom_point(size=3, alpha = 0.5)+geom_smooth(method=lm, se= F)
```

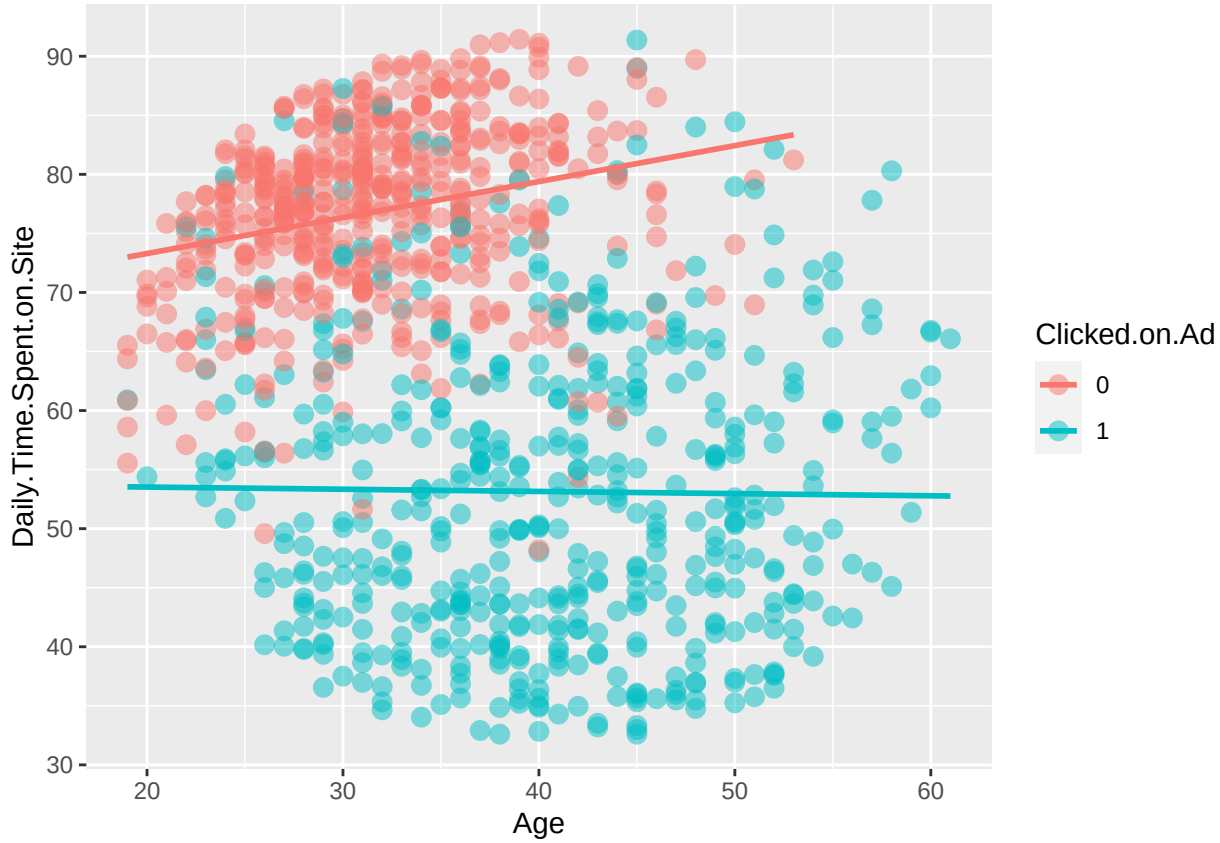
```
## `geom_smooth()` using formula 'y ~ x'
```



Majority of those who clicked on ad through Jan to July spent less time online

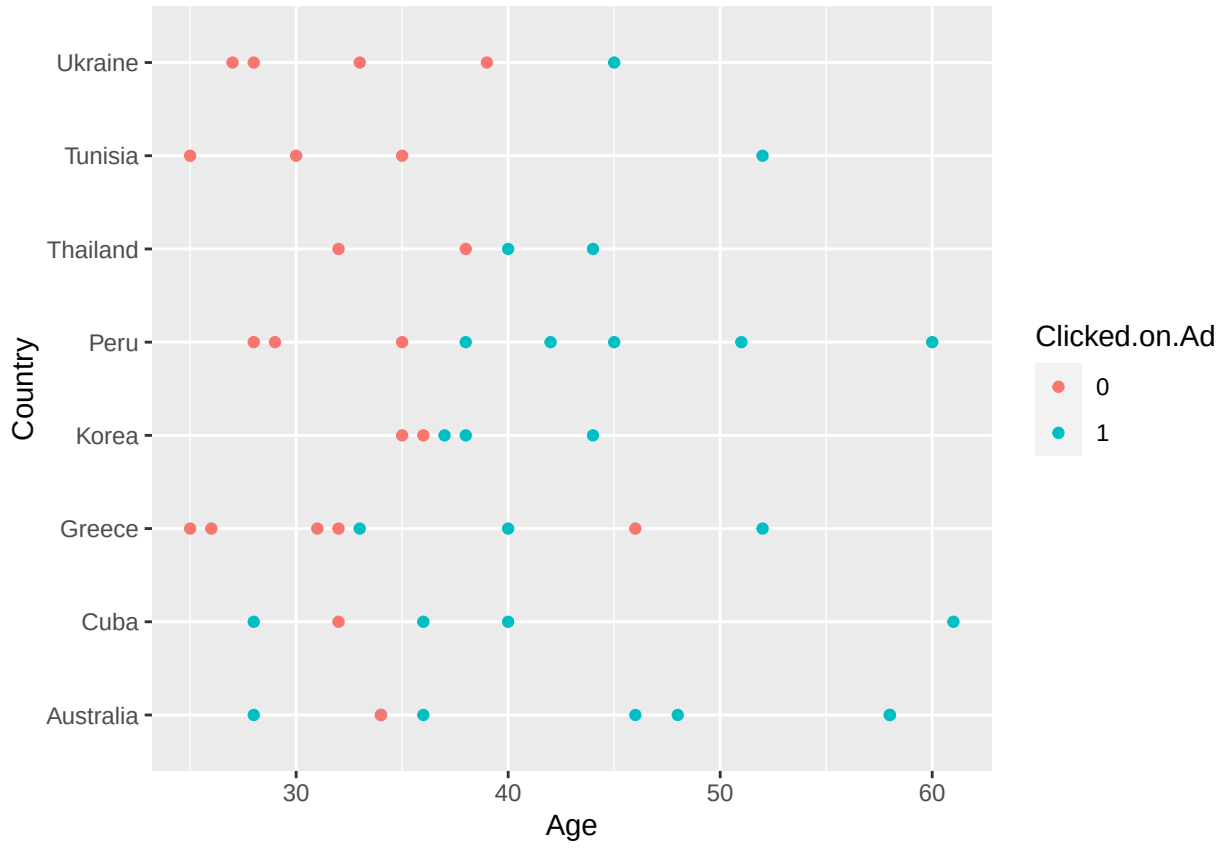
```
# A scatter plot of Age VS time spent on site
#
ads %>% ggplot(aes(Age,Daily.Time.Spent.on.Site, colour = Clicked.on.Ad))+
  geom_point(size=3, alpha = 0.5)+geom_smooth(method=lm, se= F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Most people who clicked on ads are aged between 30 -50 years and they spent much less time online

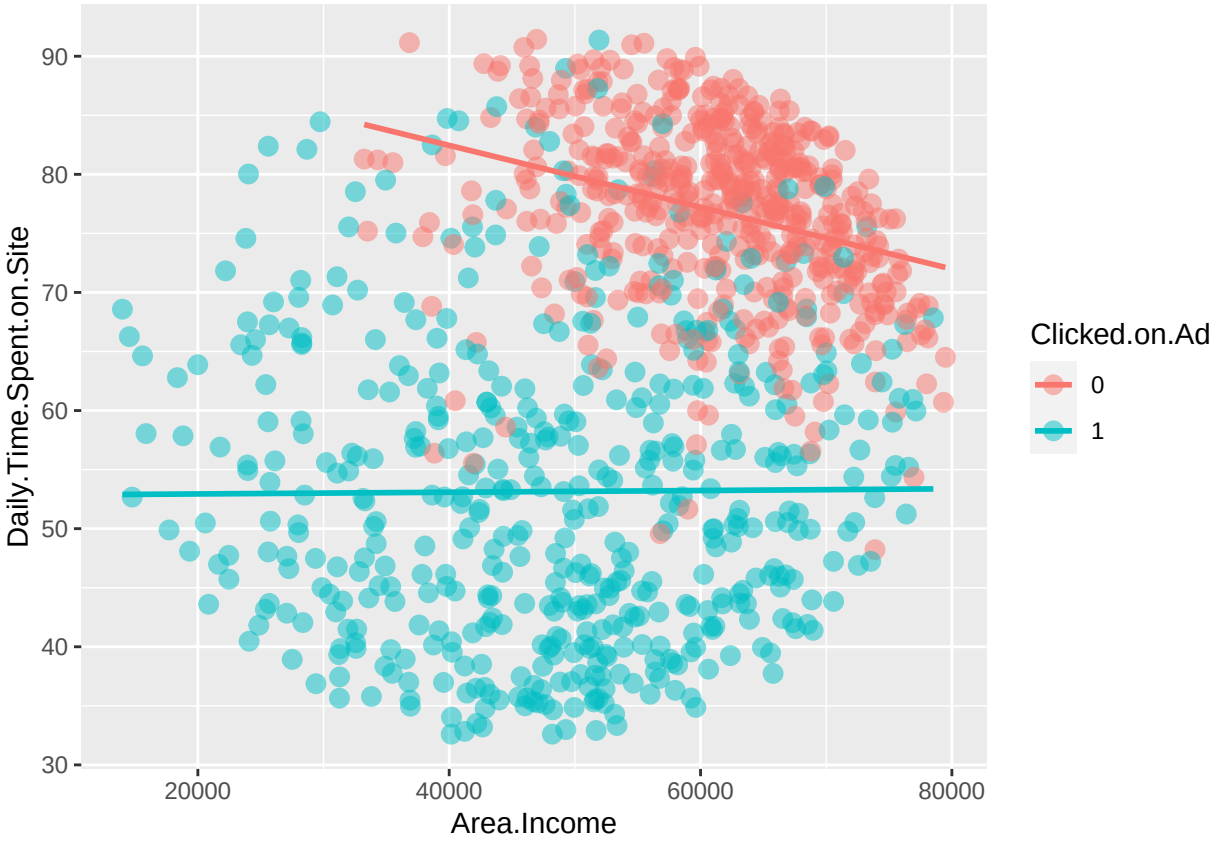
```
# A scatter plot of random selected countries VS Ages
#
filter(ads, Country %in% c("Cuba","Tunisia","Korea","Peru","Thailand","Greece","Senrgal","Ukraine","Aus
  ggplot(aes(Age,Country , colour=Clicked.on.Ad)) +
  geom_point()
```



Most Countries the people who click on ads are 30yrs and above

```
# A scatter plot of Age VS time spent on site
#
ads %>% ggplot(aes(Area.Income,Daily.Time.Spent.on.Site, colour = Clicked.on.Ad))+
  geom_point(size=3, alpha = 0.5)+geom_smooth(method=lm, se= F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
Num <- ads %>% select(Daily.Time.Spent.on.Site, Age, Area.Income, Daily.Internet.Usage)

corr <- cor(Num)

corr
```

```
##                           Daily.Time.Spent.on.Site        Age Area.Income
## Daily.Time.Spent.on.Site                 1.0000000 -0.3315133   0.3109544
## Age                                     -0.3315133  1.0000000  -0.1826050
## Area.Income                              0.3109544 -0.1826050   1.0000000
## Daily.Internet.Usage                     0.5186585 -0.3672086   0.3374955
##                           Daily.Internet.Usage
## Daily.Time.Spent.on.Site            0.5186585
## Age                                -0.3672086
## Area.Income                         0.3374955
## Daily.Internet.Usage                1.0000000
```

Daily internet usage is positively corr to daily time spent online 0.52 , Age is negatively corr to time spent online -0.37

Modelling

```
str(ads)
```

```
## 'data.frame':    1000 obs. of  11 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 standardi
##  $ 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               : chr  "2016-03-27 00:53:11" "2016-04-04 01:39:02" "2016-03-13 20:35:42"
##  $ Clicked.on.Ad           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
##  $ Date                    : Date, format: "2016-03-27" "2016-04-04" ...
```

```
attach(ads)
```

```
## The following objects are masked from ads (pos = 3):
##
##     Ad.Topic.Line, Age, Area.Income, City, Clicked.on.Ad, Country,
##     Daily.Internet.Usage, Daily.Time.Spent.on.Site, Date, Male,
##     Timestamp
```

```
ads$Male <- as.numeric(Male)
ads$Clicked.on.Ad<- as.numeric(Clicked.on.Ad)
ads$Age<- as.numeric(Age)
```

```
data <- subset(ads, select = -c(Timestamp,Country,City,Ad.Topic.Line,Date))
 head(data)
```

```
##   Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage Male
## 1                    68.95  35    61833.90               256.09    1
## 2                    80.23  31    68441.85               193.77    2
## 3                    69.47  26    59785.94               236.50    1
## 4                    74.15  29    54806.18               245.89    2
## 5                    68.37  35    73889.99               225.58    1
## 6                    59.99  23    59761.56               226.74    2
##   Clicked.on.Ad
## 1             1
## 2             1
## 3             1
## 4             1
## 5             1
## 6             1
```

```
glimpse(data)
```

```
## Rows: 1,000
## Columns: 6
## $ Daily.Time.Spent.on.Site <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, 88.~
```

```
## $ Age                       <dbl> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49, 3~
## $ Area.Income               <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 73889~
## $ Daily.Internet.Usage      <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 226.7~
## $ Male                      <dbl> 1, 2, 1, 2, 1, 2, 1, 2, 2, 2, 1, 2, 2, 1, 1, ~
## $ Clicked.on.Ad             <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 1, 2, 1, 2, ~
```

```
## KNN model
# performing normalization
normal <- function(x) (
  return( ((x - min(x)) /(max(x)-min(x))) )
)

ads_norm <- as.data.frame(lapply(data[1:5], normal))
head(ads_norm)
```

```
##   Daily.Time.Spent.on.Site       Age Area.Income Daily.Internet.Usage Male
## 1                0.6178820 0.3809524   0.7304725            0.9160310    0
## 2                0.8096209 0.2857143   0.8313752            0.5387456    1
## 3                0.6267211 0.1666667   0.6992003            0.7974331    0
## 4                0.7062723 0.2380952   0.6231599            0.8542802    1
## 5                0.6080231 0.3809524   0.9145678            0.7313234    0
## 6                0.4655788 0.0952381   0.6988280            0.7383460    1
```

```
set.seed(55) # to get same random sample

# selecting 70% of the data
ads_samp <- sample(1:nrow(ads_norm), size = nrow(ads_norm)*0.7,replace=FALSE)
```

```
# getting 70% train and 30 % test data X(Independent)
X_train<- data[ads_samp,]
X_test <- data[-ads_samp,]
```

```
# Creating train and test dataset for y (dependent)
y_train <- data[ads_samp,6]
y_test <- data[-ads_samp,6]
```

```
library(class)
# Applying k-NN classification algorithm.
# No. of neighbours are generally square root of total number of instances
neigh<- round(sqrt(NROW(y_train)))+1 # here we want to have the number y_training data
# Applying the knn algorithm
model<- knn(train = X_train,  test = X_test, cl = y_train, k = neigh)
```

```
# getting a confussion matrix
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
confusionMatrix(table(y_test, model))
```

```
## Confusion Matrix and Statistics
##
##       model
## y_test   1   2
##      1 114  32
##      2  60  94
##
##                Accuracy : 0.6933
##                  95% CI : (0.6378, 0.745)
##     No Information Rate : 0.58
##     P-Value [Acc > NIR] : 3.413e-05
##
##                   Kappa : 0.3893
##
##  Mcnemar's Test P-Value : 0.004879
##
##             Sensitivity : 0.6552
##             Specificity : 0.7460
##          Pos Pred Value : 0.7808
##          Neg Pred Value : 0.6104
##              Prevalence : 0.5800
##          Detection Rate : 0.3800
##    Detection Prevalence : 0.4867
##       Balanced Accuracy : 0.7006
##
##        'Positive' Class : 1
##
```

```
# Calculating the Accuracy
mean(y_test== model)*100
```

```
## [1] 69.33333
```

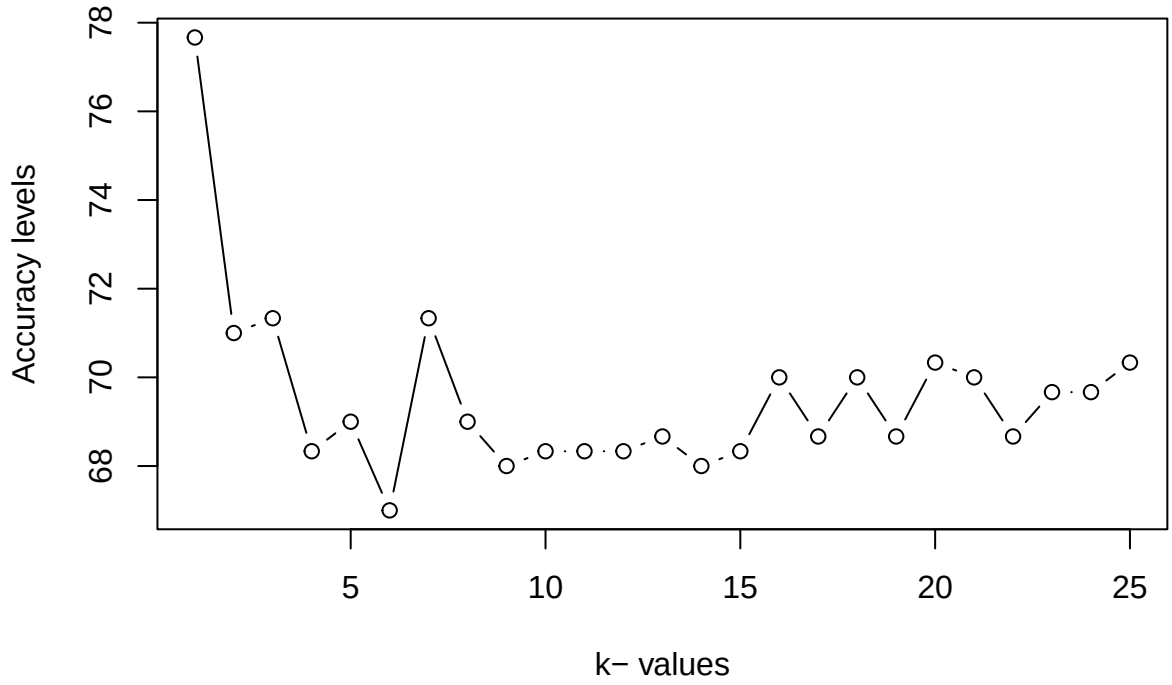
```
# Tuning the model
i = 1  #initiating a loop
k.optm = 1
for (i in 1:25) {
  model<- knn(train = X_train,  test = X_test, cl=y_train, k=i)
  k.optm[i] <- mean(y_test== model)*100
  k=i
  cat(k ,"=", k.optm[i],'\n')   # to print accuracy

}
```

```
## 1 = 77.66667
## 2 = 71
```

```
## 3 = 71.33333
## 4 = 68.33333
## 5 = 69
## 6 = 67
## 7 = 71.33333
## 8 = 69
## 9 = 68
## 10 = 68.33333
## 11 = 68.33333
## 12 = 68.33333
## 13 = 68.66667
## 14 = 68
## 15 = 68.33333
## 16 = 70
## 17 = 68.66667
## 18 = 70
## 19 = 68.66667
## 20 = 70.33333
## 21 = 70
## 22 = 68.66667
## 23 = 69.66667
## 24 = 69.66667
## 25 = 70.33333
```

```
plot(k.optm, type="b", xlab="k- values", ylab="Accuracy levels") #to print optimum accuracy level
```



* The model performs better at K = 1 with a 77.67 % accuracy.

Conclusion

Ads are mostly clicked: Those aged between 30 and 60 years Those who spent less time online Ads are clicked through out the months.

Recommendations

Its Recommend that to target age group between 30 and 60, have a well detailed and well explained advert as those who click on ads do spent much time on site and make sure advert run through out the year