Supervised IP

Daisy Lynn

2022-03-18

Defining the Question

A. Specifying the Data Analytic Question

Identify which individuals are most likely to click on ads

B. Defining the Metric for Success

Research will be considered a success when data is throughly cleaned and relationship/ effect of variables on target variable ‘Click on Ad’ is determined

c. Understanding 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

D. Experimental Design

Reading the data Checking the data Cleaning dataset Univariate Analysis Bivariate Analysis Conclusions

E. Data Relevance

link to dataset http://bit.ly/IPAdvertisingData

2. Reading the Data

```
library(readr)
df <- read_csv("http://bit.ly/IPAdvertisingData")
```

```
## Rows: 1000 Columns: 10
## -- Column specification ------------------------------------------------------
## Delimiter: ","
## chr  (3): Ad Topic Line, City, Country
## dbl  (6): Daily Time Spent on Site, Age, Area Income, Daily Internet Usage, ...
## dttm (1): Timestamp
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
head(df)
```

```
## # A tibble: 6 x 10
##   ‘Daily Time Spent~‘   Age ‘Area Income‘ ‘Daily Interne~‘ ‘Ad Topic Line‘ City
##                 <dbl> <dbl>         <dbl>            <dbl> <chr>           <chr>
## 1                69.0    35        61834.             256. Cloned 5thgene~ Wrig~
## 2                80.2    31        68442.             194. Monitored nati~ West~
## 3                69.5    26        59786.             236. Organic bottom~ Davi~
## 4                74.2    29        54806.             246. Triple-buffere~ West~
## 5                68.4    35        73890.             226. Robust logisti~ Sout~
## 6                60.0    23        59762.             227. Sharable clien~ Jami~
## # ... with 4 more variables: Male <dbl>, Country <chr>, Timestamp <dttm>,
## #   ‘Clicked on Ad‘ <dbl>
```

```
library(readr)
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ggplot2)
library(tidyverse)
```

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

## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v purrr   0.3.4     v forcats 0.5.1

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

```
library(dplyr)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(e1071)
library(cluster)
library(kernlab)
```

```
## 
## Attaching package: 'kernlab'

## The following object is masked from 'package:purrr':
## 
##     cross

## The following object is masked from 'package:ggplot2':
## 
##     alpha
```

```
library(tidyr)
library(tinytex)
library(superml)
```

```
## Loading required package: R6
```

3. Checking the Data

```
# tail of dataset

head(df)
```

```
## # A tibble: 6 x 10
##   ‘Daily Time Spent~‘   Age ‘Area Income‘ ‘Daily Interne~‘ ‘Ad Topic Line‘ City 
##                 <dbl> <dbl>         <dbl>            <dbl> <chr>           <chr>
## 1                69.0    35        61834.             256. Cloned 5thgene~ Wrig~
## 2                80.2    31        68442.             194. Monitored nati~ West~
## 3                69.5    26        59786.             236. Organic bottom~ Davi~
## 4                74.2    29        54806.             246. Triple-buffere~ West~
## 5                68.4    35        73890.             226. Robust logisti~ Sout~
## 6                60.0    23        59762.             227. Sharable clien~ Jami~
## # ... with 4 more variables: Male <dbl>, Country <chr>, Timestamp <dttm>,
## #   ‘Clicked on Ad‘ <dbl>
```

```
# tail of dataset

tail(df)
```

```
## # A tibble: 6 x 10
##   ‘Daily Time Spent~‘   Age ‘Area Income‘ ‘Daily Interne~‘ ‘Ad Topic Line‘ City 
##                 <dbl> <dbl>         <dbl>            <dbl> <chr>           <chr>
## 1                43.7    28        63127.             173. Front-line bif~ Nich~
## 2                73.0    30        71385.             209. Fundamental mo~ Duff~
## 3                51.3    45        67782.             134. Grass-roots co~ New ~
## 4                51.6    51        42416.             120. Expanded intan~ Sout~
## 5                55.6    19        41921.             188. Proactive band~ West~
## 6                45.0    26        29876.             178. Virtual 5thgen~ Ronn~
## # ... with 4 more variables: Male <dbl>, Country <chr>, Timestamp <dttm>,
## #   ‘Clicked on Ad‘ <dbl>
```

```
#checking size of dataframe

dim(df)
```

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

The dataset has 1000 rows and 10 variables

```
#cheking columns

colnames(df)
```

```
##  [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"
```

4. Cleaning dataset

```
# Identifying missing data in dataset

colSums(is.na(df))
```

```
## 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 presence of null values

```
#finding null values

is.null(df)
```

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

```
#checking unique values of age variable

unique(df$Age)
```

```
##  [1] 35 31 26 29 23 33 48 30 20 49 37 24 41 36 40 52 28 34 22 57 53 39 46 32 25
## [26] 43 45 50 47 27 42 38 54 21 60 55 44 58 56 51 19 59 61
```

```
#checking unique values of Male variable

unique(df$Male)
```

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

Above results show if individual is male or not 0 represents no and 1 represents yes

```
#checking unique values of Clicked on Ad variable

unique(df$`Clicked on Ad`)
```

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

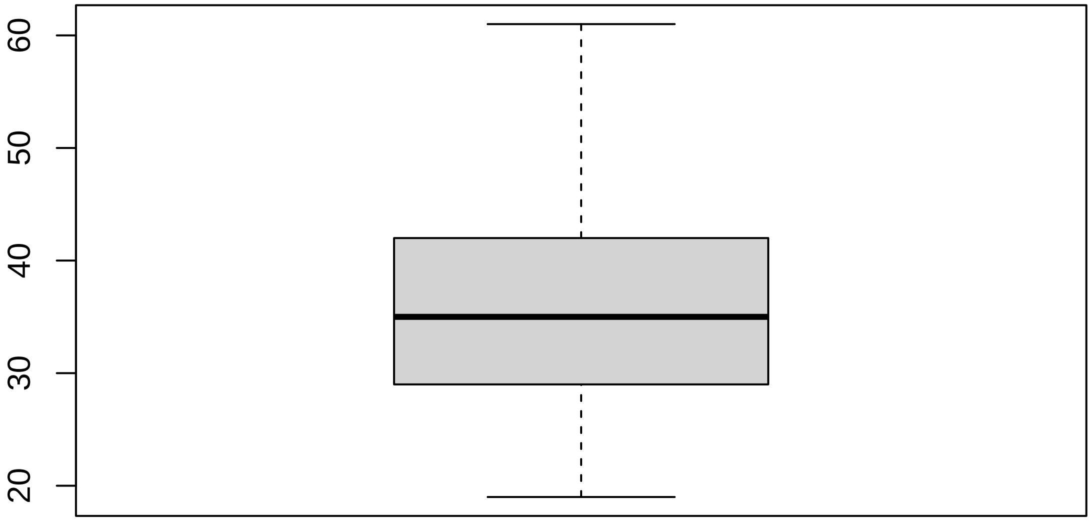
checking if audience clicked on Ad or not 0 means no 1 means yes

```
#checking unique values of Country variable

unique(df$Country)
```

```
##   [1] "Tunisia"
##   [2] "Nauru"
##   [3] "San Marino"
##   [4] "Italy"
##   [5] "Iceland"
##   [6] "Norway"
##   [7] "Myanmar"
##   [8] "Australia"
##   [9] "Grenada"
##  [10] "Ghana"
##  [11] "Qatar"
##  [12] "Burundi"
##  [13] "Egypt"
##  [14] "Bosnia and Herzegovina"
##  [15] "Barbados"
##  [16] "Spain"
##  [17] "Palestinian Territory"
##  [18] "Afghanistan"
##  [19] "British Indian Ocean Territory (Chagos Archipelago)"
##  [20] "Russian Federation"
##  [21] "Cameroon"
##  [22] "Korea"
##  [23] "Tokelau"
##  [24] "Monaco"
##  [25] "Tuvalu"
##  [26] "Greece"
##  [27] "British Virgin Islands"
##  [28] "Bouvet Island (Bouvetoya)"
##  [29] "Peru"
##  [30] "Aruba"
##  [31] "Maldives"
##  [32] "Senegal"
##  [33] "Dominica"
##  [34] "Luxembourg"
##  [35] "Montenegro"
##  [36] "Ukraine"
##  [37] "Saint Helena"
##  [38] "Liberia"
##  [39] "Turkmenistan"
##  [40] "Niger"
##  [41] "Sri Lanka"
##  [42] "Trinidad and Tobago"
##  [43] "United Kingdom"
##  [44] "Guinea-Bissau"
##  [45] "Micronesia"
##  [46] "Turkey"
##  [47] "Croatia"
##  [48] "Israel"
##  [49] "Svalbard & Jan Mayen Islands"
##  [50] "Azerbaijan"
##  [51] "Iran"
##  [52] "Saint Vincent and the Grenadines"
##  [53] "Bulgaria"
##  [54] "Christmas Island"
##  [55] "Canada"
##  [56] "Rwanda"
##  [57] "Turks and Caicos Islands"
##  [58] "Norfolk Island"
##  [59] "Cook Islands"
##  [60] "Guatemala"
##  [61] "Cote d'Ivoire"
##  [62] "Faroe Islands"
##  [63] "Ireland"
##  [64] "Moldova"
##  [65] "Nicaragua"
##  [66] "Montserrat"
##  [67] "Timor-Leste"
##  [68] "Puerto Rico"
##  [69] "Central African Republic"
##  [70] "Venezuela"
##  [71] "Wallis and Futuna"
##  [72] "Jersey"
##  [73] "Samoa"
##  [74] "Antarctica (the territory South of 60 deg S)"
##  [75] "Albania"
##  [76] "Hong Kong"
##  [77] "Lithuania"
##  [78] "Bangladesh"
##  [79] "Western Sahara"
##  [80] "Serbia"
##  [81] "Czech Republic"
##  [82] "Guernsey"
##  [83] "Tanzania"
##  [84] "Bhutan"
##  [85] "Guinea"
##  [86] "Madagascar"
##  [87] "Lebanon"
##  [88] "Eritrea"
##  [89] "Guyana"
##  [90] "United Arab Emirates"
##  [91] "Martinique"
##  [92] "Somalia"
##  [93] "Benin"
##  [94] "Papua New Guinea"
##  [95] "Uzbekistan"
##  [96] "South Africa"
##  [97] "Hungary"
##  [98] "Falkland Islands (Malvinas)"
##  [99] "Saint Martin"
## [100] "Cuba"
## [101] "United States Minor Outlying Islands"
## [102] "Belize"
## [103] "Kuwait"
## [104] "Thailand"
## [105] "Gibraltar"
## [106] "Holy See (Vatican City State)"
## [107] "Netherlands"
## [108] "Belarus"
## [109] "New Zealand"
## [110] "Togo"
## [111] "Kenya"
## [112] "Palau"
## [113] "Cambodia"
## [114] "Costa Rica"
## [115] "Liechtenstein"
## [116] "Angola"
## [117] "Equatorial Guinea"
## [118] "Mongolia"
## [119] "Brazil"
## [120] "Chad"
## [121] "Portugal"
## [122] "Malawi"
## [123] "Singapore"
## [124] "Kazakhstan"
## [125] "China"
## [126] "Vietnam"
## [127] "Mayotte"
## [128] "Jamaica"
## [129] "Bahamas"
## [130] "Algeria"
## [131] "Fiji"
## [132] "Argentina"
## [133] "Philippines"
## [134] "Suriname"
## [135] "Guam"
## [136] "Antigua and Barbuda"
## [137] "Georgia"
## [138] "Jordan"
## [139] "Saudi Arabia"
## [140] "Sao Tome and Principe"
## [141] "Cyprus"
## [142] "Kyrgyz Republic"
## [143] "Pakistan"
## [144] "Seychelles"
## [145] "Mauritania"
## [146] "Chile"
## [147] "Poland"
## [148] "Estonia"
## [149] "Latvia"
## [150] "Bahrain"
## [151] "Colombia"
## [152] "Brunei Darussalam"
## [153] "Taiwan"
## [154] "Saint Pierre and Miquelon"
## [155] "Finland"
## [156] "French Southern Territories"
## [157] "Sierra Leone"
## [158] "Tajikistan"
## [159] "Ecuador"
## [160] "Switzerland"
## [161] "France"
## [162] "Malaysia"
## [163] "Mauritius"
## [164] "Japan"
## [165] "Greenland"
## [166] "Guadeloupe"
## [167] "Belgium"
## [168] "Honduras"
## [169] "Paraguay"
## [170] "French Guiana"
## [171] "Northern Mariana Islands"
## [172] "American Samoa"
## [173] "Austria"
## [174] "Tonga"
## [175] "New Caledonia"
## [176] "United States of America"
## [177] "Morocco"
## [178] "Macedonia"
## [179] "Gabon"
## [180] "Uganda"
## [181] "Saint Lucia"
## [182] "Niue"
## [183] "Zambia"
## [184] "Congo"
## [185] "Pitcairn Islands"
## [186] "Anguilla"
## [187] "Sweden"
## [188] "Indonesia"
## [189] "Mexico"
## [190] "Haiti"
## [191] "Gambia"
## [192] "El Salvador"
## [193] "Libyan Arab Jamahiriya"
## [194] "Saint Barthelemy"
## [195] "Reunion"
## [196] "Panama"
## [197] "Dominican Republic"
## [198] "Zimbabwe"
## [199] "Swaziland"
## [200] "Saint Kitts and Nevis"
## [201] "Burkina Faso"
## [202] "Heard Island and McDonald Islands"
## [203] "Bolivia"
## [204] "Netherlands Antilles"
## [205] "French Polynesia"
## [206] "Germany"
## [207] "Malta"
## [208] "Sudan"
## [209] "Lao People's Democratic Republic"
## [210] "Isle of Man"
## [211] "Macao"
## [212] "United States Virgin Islands"
## [213] "Djibouti"
## [214] "Mali"
## [215] "Romania"
## [216] "Cayman Islands"
## [217] "Ethiopia"
## [218] "Uruguay"
## [219] "Comoros"
## [220] "Vanuatu"
## [221] "Nepal"
## [222] "Yemen"
## [223] "India"
## [224] "Cape Verde"
## [225] "Slovenia"
## [226] "Denmark"
## [227] "Syrian Arab Republic"
## [228] "Andorra"
## [229] "Namibia"
## [230] "Slovakia (Slovak Republic)"
## [231] "Armenia"
## [232] "South Georgia and the South Sandwich Islands"
## [233] "Kiribati"
## [234] "Marshall Islands"
## [235] "Bermuda"
## [236] "Mozambique"
## [237] "Lesotho"
```

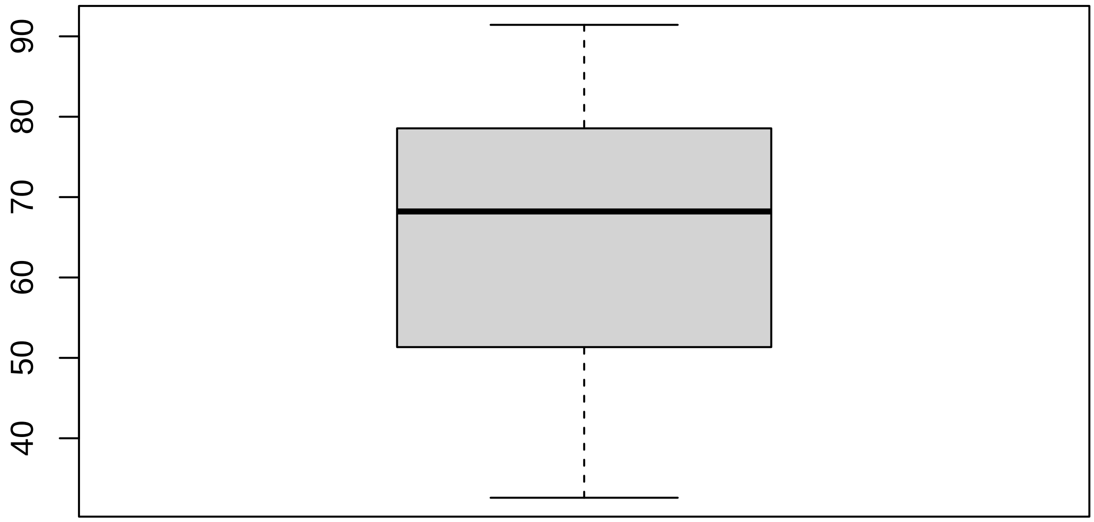
```
#finding outliers in the Age column
```

```
boxplot(df$Age)
```



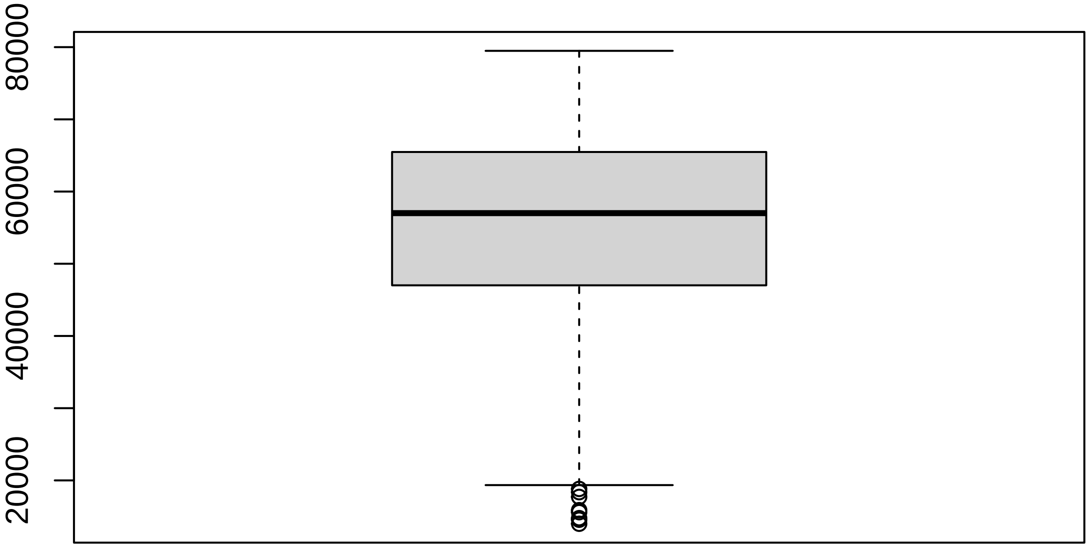
No presence of outliers

```
#finding outliers in the Daily Time Spent on Site column

boxplot(df$`Daily Time Spent on Site`)
```



No presence of outliers

```
#finding outliers in the Area Income column

boxplot(df$`Area Income`)
```



There is presence of outliers but they won't be removed since the represent real data

```
#finding duplicates

duplicates <- df[duplicated(df),]
duplicates
```

```
## # A tibble: 0 x 10
## # ... with 10 variables: Daily Time Spent on Site <dbl>, Age <dbl>,
## #   Area Income <dbl>, Daily Internet Usage <dbl>, Ad Topic Line <chr>,
## #   City <chr>, Male <dbl>, Country <chr>, Timestamp <dttm>,
## #   Clicked on Ad <dbl>
```

No presence of duplicates

5. Univariate Analysis

```
# getting the summary of our numerical columns

summary(df)
```

```
##  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
```

```
# finding range of Age variable
Age.range <- range(df$Age)
Age.range
```

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

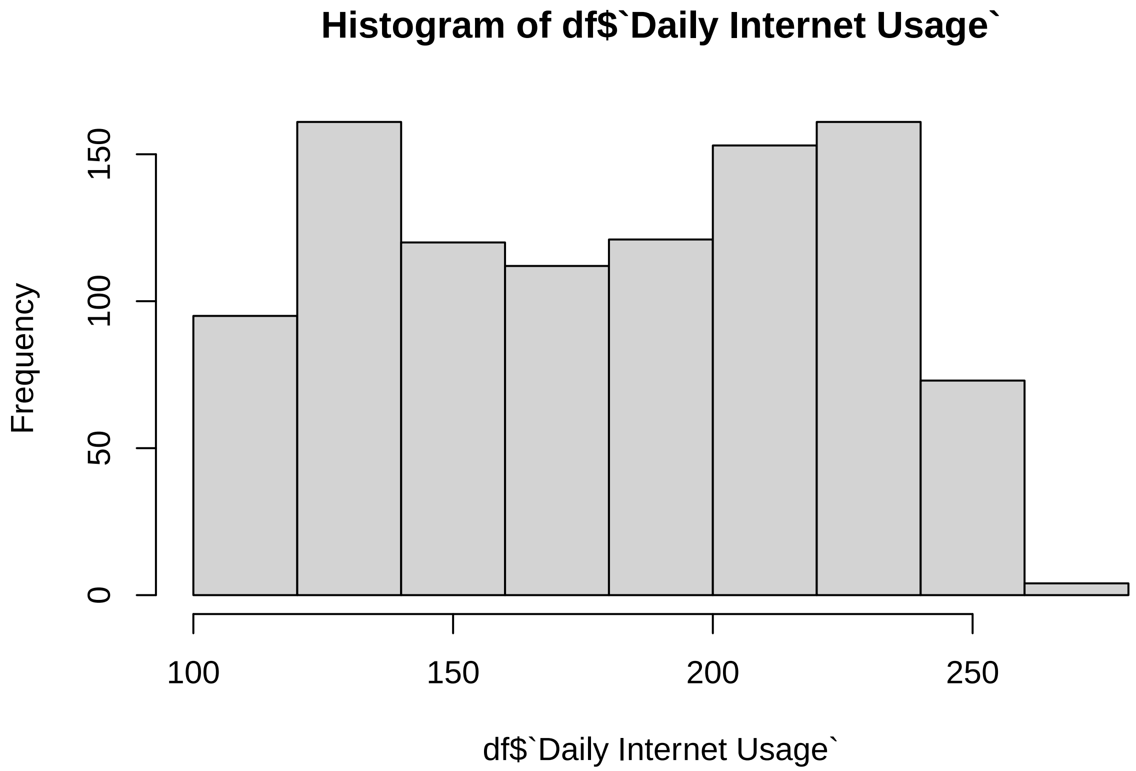
```
# finding range of Daily Time Spent on Site variable
TimeSpent.range <- range(df$`Daily Time Spent on Site` )
TimeSpent.range
```

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

```
#Variance of Area Income variable
AreaIncome.variance <- var(df$`Area Income`)
#
AreaIncome.variance
```

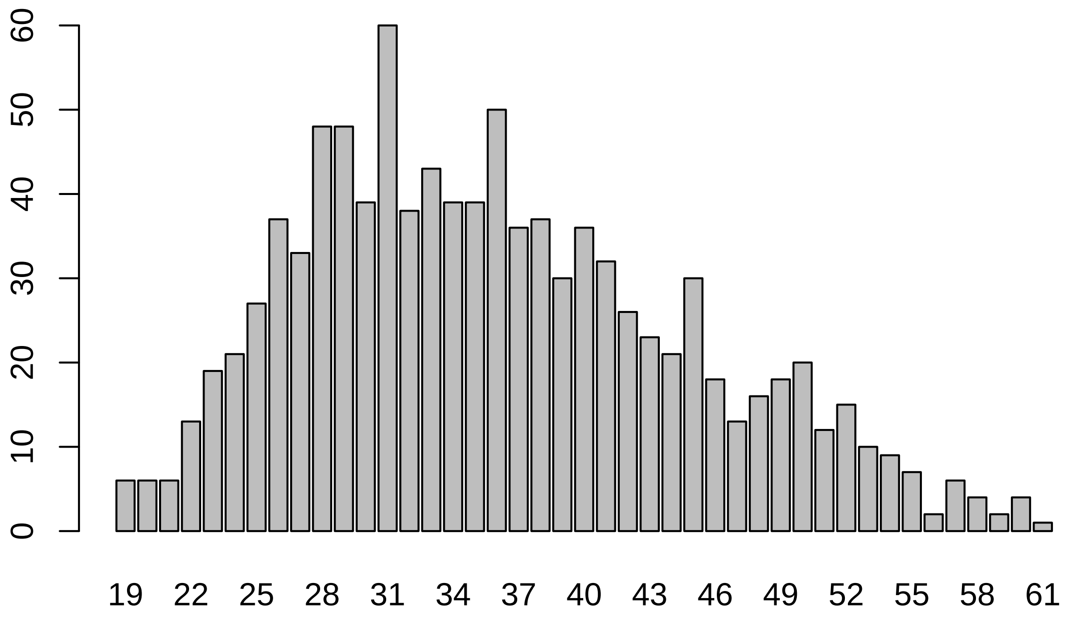
```
## [1] 179952406
```

```
#Histogram visualization on Daily Internet Usage

hist(df$`Daily Internet Usage`)
```



From above visualization most of the daily internet usage was between 200 to 250

```r
#Barplot visualization for Age variable

Age <- df$ Age
Age_frequency <- table(Age)
barplot(Age_frequency)
```



From above visualization most of the audience are 31 yrs of age

6. Bivariate and Multivariate Analysis

```
#Covariance between Daily Time Spent on Site and Daily Internet Usage

TimeSpentonSite <- df$`Daily Time Spent on Site`
#
InternetUsage<- df$`Daily Internet Usage`

#
cov(TimeSpentonSite,InternetUsage)
```

```
## [1] 360.9919
```

The result is positive, meaning that the variables are positively related.

```
#Covariance between age and Area Income

Age <- df$Age
#
Income<- df$`Area Income`
```

```r
#
cov(Age, Income)
```

```
## [1] -21520.93
```

The result is negative, meaning that the variables are negatively related.

```r
#Correlation between age and clicked on ad

Age <- df$Age
#
Ad<- df$`Clicked on Ad`

#
cor(Age, Ad)
```

```
## [1] 0.4925313
```

There is a moderate positive correlation between the two variables

```r
#Correlation between Male and clicked on ad

Male <- df$Male
#
Ad<- df$`Clicked on Ad`

#
cor(Male, Ad)
```

```
## [1] -0.03802747
```

Weak negative correlation

```r
#Correlation between Ad and Internet Usage

InternetUsage<- df$`Daily Internet Usage`
#
Ad<- df$`Clicked on Ad`

#
cor(InternetUsage, Ad)
```

```
## [1] -0.7865392
```

Strong negative relation

```r
#Correlation between Ad and Timespent on site

TimeSpentonSite <- df$`Daily Time Spent on Site`
#
```

```
Ad<- df$`Clicked on Ad`

#
cor(TimeSpentonSite, Ad)
```

```
## [1] -0.7481166
```

Strong negative relationship

```
#Correlation between Male and Daily Internet Usage

Male <- df$Male
#
InternetUsage<- df$`Daily Internet Usage`

#
cor(Male, InternetUsage)
```

```
## [1] 0.02801233
```

There is a positive correlation between the above variables although very weak

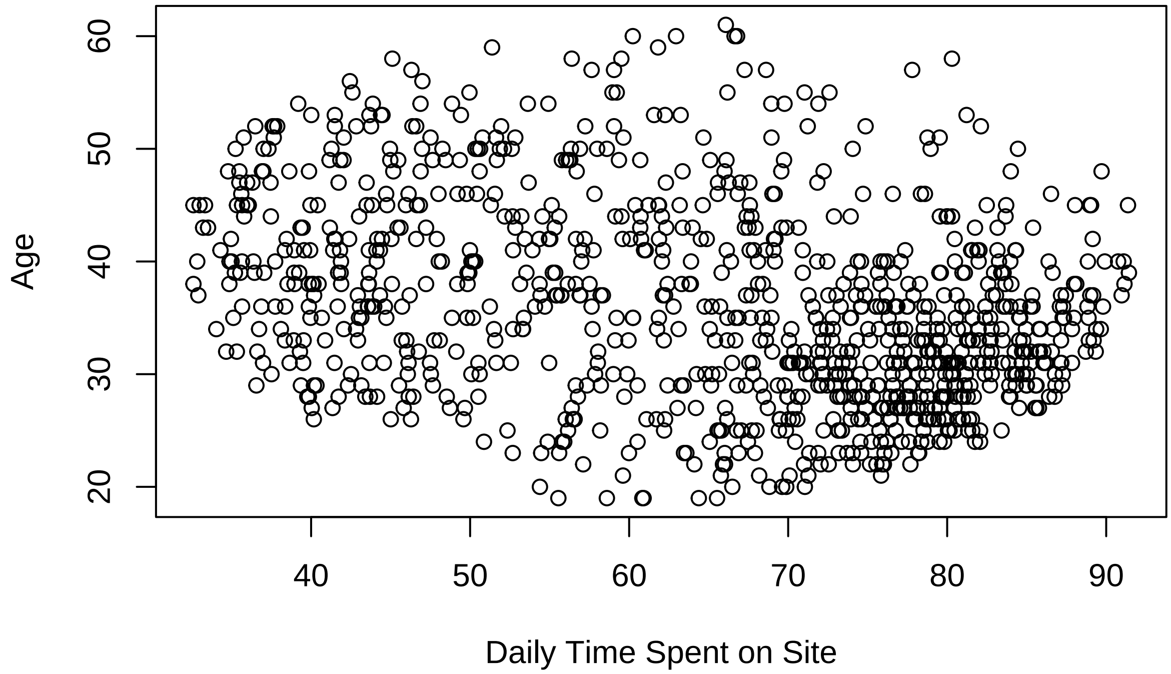
```
#Correlation between age and Area Income

Age <- df$Age
#
Income<- df$`Area Income`

#
cor(Age, Income)
```

```
## [1] -0.182605
```

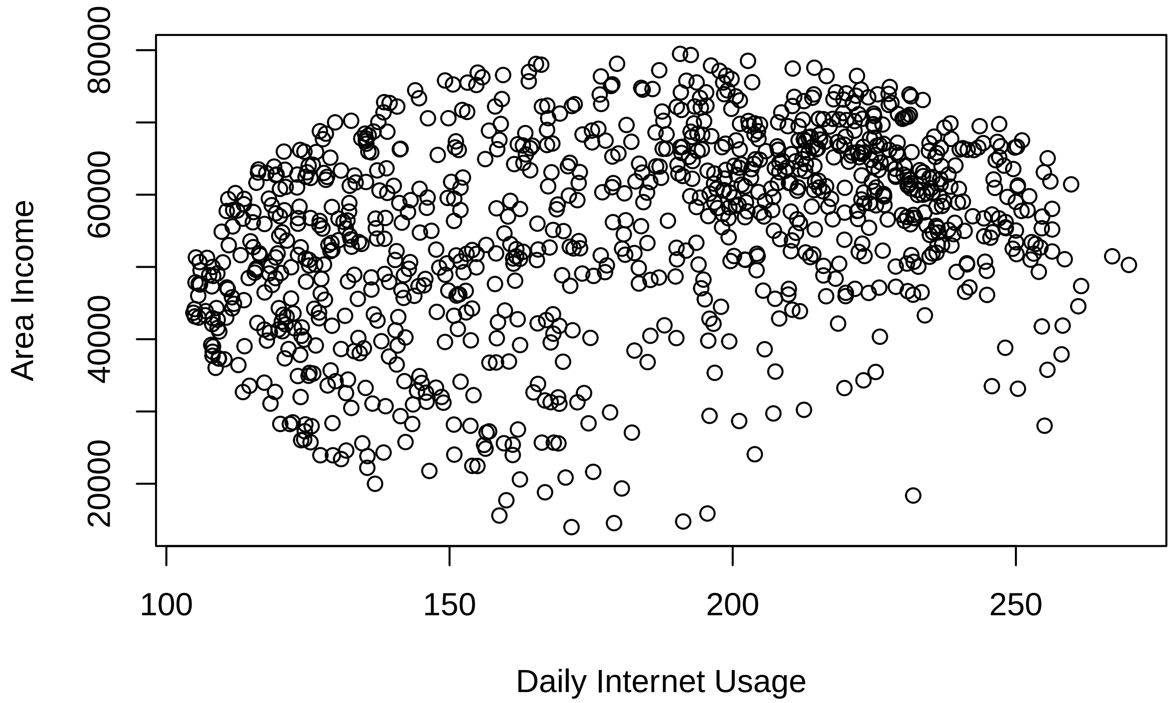
There is a weak negative correlation between the above variables

6.1 Visualization

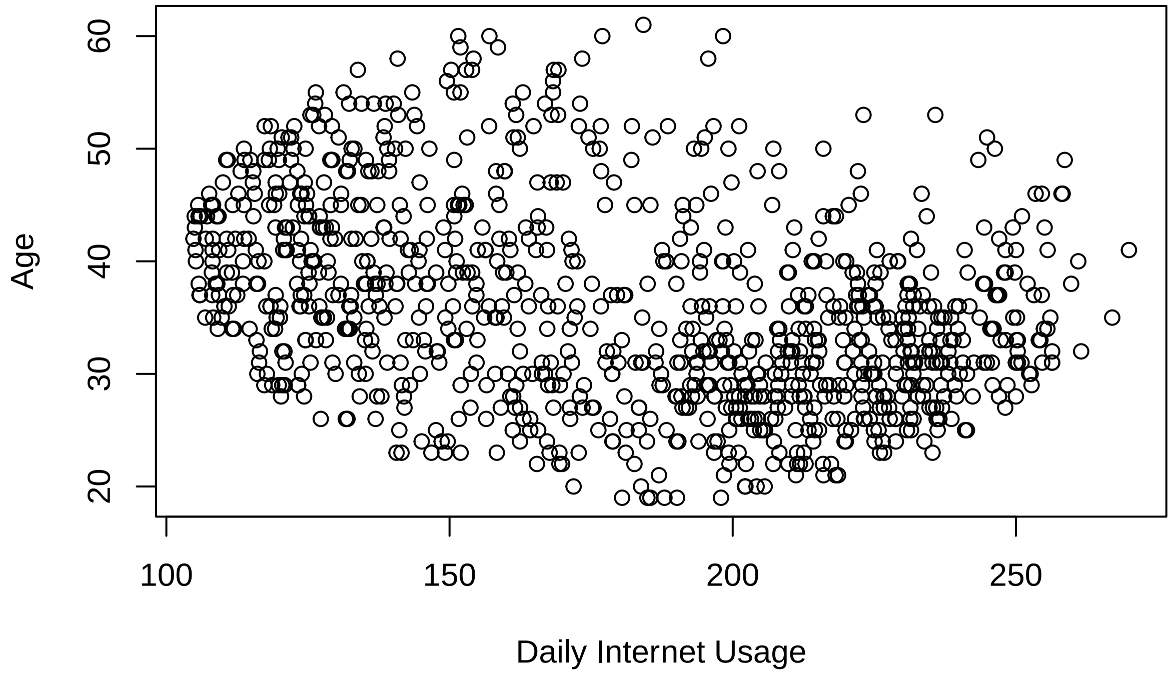
```
#Scatter plot on the relation between Daily Time Spent on Site and Age

TimeSpentonSite <- df$`Daily Time Spent on Site`
#
Age<- df$`Age`

#
plot(TimeSpentonSite, Age, xlab="Daily Time Spent on Site", ylab="Age")
```



```
#Scatter plot on the relation between Area Income and Daily Internet Usage

Income <- df$`Area Income`
#
InternetUsage<- df$`Daily Internet Usage`

#
plot(InternetUsage, Income, xlab="Daily Internet Usage", ylab="Area Income")
```



```
#Scatter plot on the relation between Daily Internet usage and Age

Internet <- df$`Daily Internet Usage`
#
Age<- df$`Age`

#
plot(Internet, Age, xlab="Daily Internet Usage", ylab="Age")
```



7. Implementing solution

7.1 Encoding, Splitting and normalization

```
lbl <- LabelEncoder$new()


df <- df %>%
  mutate(City = factor(lbl$fit_transform(.$City)),
         Country = factor(lbl$fit_transform(.$Country)),
         Male = factor(.$Male),
         `Clicked on Ad`= factor(.$`Clicked on Ad`),) %>%
  select(-"Timestamp",-"Ad Topic Line")
```

TimeStamp column and Ad topic line was dropped since it isnt that important for our prediction

```
#Splitting

intrain <- createDataPartition(y = df$`Clicked on Ad`, p= 0.8, list = FALSE)
training <- df[intrain,]
testing <- df[-intrain,]
```

```
# We check the dimensions of out training dataframe and testing dataframe
# ---
#
dim(training);
```

```
## [1] 800   8
```

```
dim(testing)
```

```
## [1] 200   8
```

```
#convert target into factor

training[["Clicked on Ad"]] = factor(training[["Clicked on Ad"]])
```

```
#normalizing dataset

normalize <- function(x){
  return ((x-min(x)) / (max(x)-min(x)))
}
```

7.2 SVM

```
svm_model <- svm(`Clicked on Ad` ~ ., data = training,
                  trControl = trainControl(method = 'repeatedcv', number=10, repeats=3))
```

```
# Prediction
predicts <- predict(svm_model, data = testing)
```

```
#confusionMatrix(table(predicts, testing$`Clicked on Ad`))
```

7.3 Naive Bayes

```
#Training

naive <- naiveBayes(`Clicked on Ad` ~ ., data = df,trControl=trainControl(method='cv',number=10))
```

```
# Model Evalution
# ---
# Predicting our testing set
#
Predict <- predict(naive, newdata = testing )
```

```
# Getting the confusion matrix to see accuracy value and other parameter values
# ---
#
#cm> confusionMatrix(Predict, testing$`Clicked on Ad` )
```

8. Conclusions

The relationship between most of our variables and the click on Ad variable is negative this means that if a variable being compared to click on ad increases the other decreases

e.g comparing Daily Internet Usage and Click on Ad, if Internet Usage increases the chance of clicking on the Ad decreases

Although for the Age variable, an increase in Age increases chance of clicking on Ad