Data Cleaning & EDA

Beaty

8/25/2021

Analysis on Ads

Data Understanding

1. Define the question:

I am a data scientist working to identify which individuals are most likely to click on my client's ads on her blog.

2. Metric for success:

- Cleaned data.
- Graphical representation of the relationships in the data as well as the distributions f the different variables in the data.
- Sound conclusions and recommendations to the client as per the analysis done.

3. 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. She would now like to employ my services as a Data Science Consultant to help her identify which individuals are most likely to click on her ads.

4. Experimental design:

Steps to be undertaken during this study include: - Loading the data & needed packages. - Exploring the dataset. - Cleaning the data. - Feature engineering. - Exploratory Data Analysis. - Conclusions. - Recommendations.

5. Data appropriateness:

This will be well checked & described in the data cleaning.

Exploring the data

```
# Loading the libraries
#install.packages("corrplot")
library("corrplot")
```

```
## corrplot 0.90 loaded
```

```r
#install.packages("PerformanceAnalytics")
library("PerformanceAnalytics")
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
##
##     legend
```

```r
#install.packages("ggplot2")
library("ggplot2")

# Loading the data
data = read.csv(url("http://bit.ly/IPAdvertisingData"))

# Previewing the data
head(data)
```

```
##   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 shape of the data
dim(data)

## [1] 1000   10
```

The dataset has 1000 entries and 10 columns.

```
# Checking column names of our data
colnames(data)

##  [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"
```

Our column titles are as listed above. We shall rename the “Male” column so that it becomes “Gender” then the 0 will represent males, and 0 for females.

```
# Renaming "Male" column
colnames(data)[7] <- "Gender"

# Checking if column name has changed
colnames(data)

##  [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"
```

Data Cleaning

1. Dealing with missing data

```
# Checking for missing values
null_values = is.na(data)
null_values

##         Daily.Time.Spent.on.Site   Age Area.Income Daily.Internet.Usage
##    [1,]                    FALSE FALSE       FALSE                 FALSE
##    [2,]                    FALSE FALSE       FALSE                 FALSE
##    [3,]                    FALSE FALSE       FALSE                 FALSE
##    [4,]                    FALSE FALSE       FALSE                 FALSE
##    [5,]                    FALSE FALSE       FALSE                 FALSE
##    [6,]                    FALSE FALSE       FALSE                 FALSE
##    [7,]                    FALSE FALSE       FALSE                 FALSE
##    [8,]                    FALSE FALSE       FALSE                 FALSE
##    [9,]                    FALSE FALSE       FALSE                 FALSE
##   [10,]                    FALSE FALSE       FALSE                 FALSE
##   [11,]                    FALSE FALSE       FALSE                 FALSE
##   [12,]                    FALSE FALSE       FALSE                 FALSE
##   [13,]                    FALSE FALSE       FALSE                 FALSE
```

```
…
##          Ad.Topic.Line  City Gender Country Timestamp Clicked.on.Ad
##    [1,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##    [2,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##    [3,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##    [4,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##    [5,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##    [6,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##    [7,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##    [8,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##    [9,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##   [10,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##   [11,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##   [12,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
##   [13,]         FALSE FALSE  FALSE   FALSE     FALSE         FALSE
…
```

Our data seems to have no nulls since most of the fields are showing “FALSE” for our code. We shall do a coun to check for the total number of null values.

```
length(which(is.na(data)))
```

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

The data has no missing values

2. Checking for duplicates:

```
duplicates = duplicated(data)
duplicates
```

```
##    [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE
##   [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE
##   [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE F
ALSE ...
```

The data seems to have no duplicates. We shall do a count just to make sure.

```
length(which(duplicated(data)))
```

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

There are no duplicates in our dataset.

3. Checking column data types

```
# Checking the columns datatypes
str(data)
```

```
## '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" "Mo
nitored national standardization" "Organic bottom-line service-desk" "Triple-
buffered reciprocal time-frame" ...
##  $ City                    : chr  "Wrightburgh" "West Jodi" "Davidton" "We
st Terrifurt" ...
##  $ Gender                  : 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" "2016-01-10 02:31:19" ...
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

All our columns have the right data type except for the time. We shall convert it to a timestamp for ease of calculation

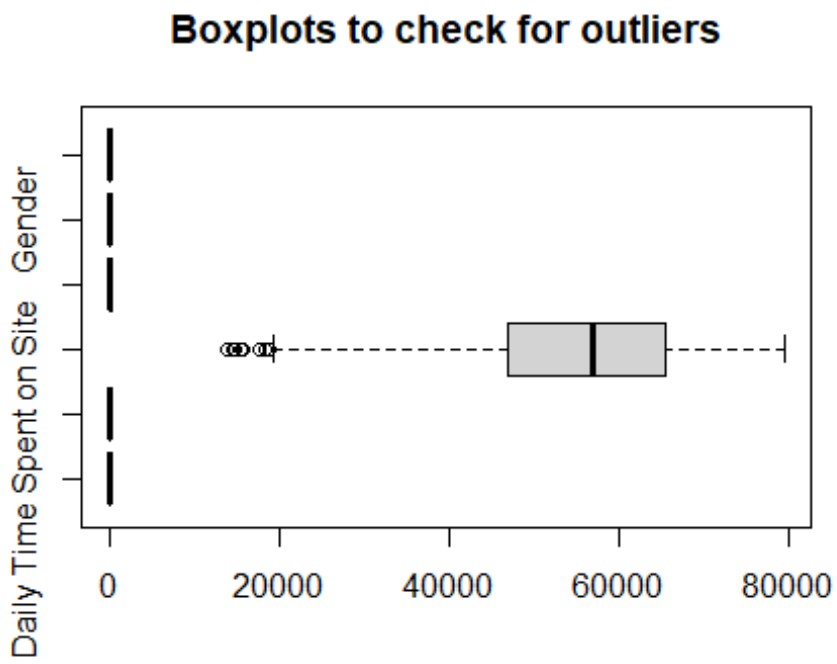
```
# Converting date from character to a timestamp
data$Timestamp <- as.Date(data$Timestamp)

# Checking data type to confirm change
str(data)

## '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" "Mo
nitored national standardization" "Organic bottom-line service-desk" "Triple-
buffered reciprocal time-frame" ...
##  $ City                    : chr  "Wrightburgh" "West Jodi" "Davidton" "We
st Terrifurt" ...
##  $ Gender                  : int  0 1 0 1 0 1 0 1 1 1 ...
##  $ Country                 : chr  "Tunisia" "Nauru" "San Marino" "Italy" .
..
##  $ Timestamp               : Date, format: "2016-03-27" "2016-04-04" ...
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

The data types are all okay now.

4. Checking for outliers

```
# Checking for outliers in the numerical columns
Time_spent = data$Daily.Time.Spent.on.Site
Age = data$Age
Income = data$Area.Income
Internet = data$Daily.Internet.Usage
Gender = data$Gender
Clicked = data$Clicked.on.Ad
```

```
boxplot(Time_spent, Age, Income, Internet, Gender, Clicked, main = "Boxplots
to check for outliers",   names = c("Daily Time Spent on Site", "Age", "Incom
e","Daily Internet Usage", "Gender","Clicked on ad"), horizontal = TRUE)
```



The columns do not have outliers except the income column which has quite a number of outliers, which may be due to the paygaps that exist in the real world. Thus we shall ignore them for this project.

Feature Engineering

We'd like to extract the month and hour from the time stamp so we can derive more insights from it.

```
data$Month <-  months(data$Timestamp)

head(data)
```

```
##   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 Gender    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   Month
## 1 2016-03-27             0   March
## 2 2016-04-04             0   April
## 3 2016-03-13             0   March
## 4 2016-01-10             0 January
## 5 2016-06-03             0    June
## 6 2016-05-19             0     May
```

Exploratory Data Analysis

1. Univariate Analysis

Measures of Dispersion

a) Mean

```
# Creating a dataframe with only the numeric columns
data_num = data[,c("Daily.Time.Spent.on.Site", "Age", "Area.Income","Daily.In
ternet.Usage" ,"Gender", "Clicked.on.Ad" )]

# Preview dataset
head(data_num)
```

```
##   Daily.Time.Spent.on.Site Age Area.Income Daily.Internet.Usage Gender
## 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
```

```
# Calculating the mean

colMeans(data_num)
```

```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##                  65.0002                  36.0090              55000.0001
##     Daily.Internet.Usage                   Gender            Clicked.on.Ad
##                180.0001                   0.4810                   0.5000
```

The means for the different columns are as shown in the output above.

b) Mode

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

# Get the mode for different columns
getmode(data$Daily.Time.Spent.on.Site)

## [1] 62.26

getmode(data$Age)

## [1] 31

getmode(data$Area.Income)

## [1] 61833.9

getmode(data$Daily.Internet.Usage)

## [1] 167.22

getmode(data$City)

## [1] "Lisamouth"

getmode(data$Gender)

## [1] 0

getmode(data$Country)

## [1] "Czech Republic"

getmode(data$Clicked.on.Ad)

## [1] 0

getmode(data$Month)

## [1] "February"
```

The most common amount of time spent on the site is 62.26, while the most popular age is 31years.

Most of the site visits have an income of 61,833.9.

The city with most visitors is Lisamouth while Czech Republic had the most site visitors. I would have expected the city with the most number of visitors to belong to the country with the most visitors, however, Lisamouth had different countries on the dataset provided.

Most of the visitors were female, while most didn't click on the ads. The site had most traffic in February.

c) Median

```
# Calculating median
median(data$Daily.Time.Spent.on.Site)

## [1] 68.215

median(data$Age)

## [1] 35

median(data$Area.Income)

## [1] 57012.3

median(data$Daily.Internet.Usage)

## [1] 183.13

median(data$Gender)

## [1] 0

median(data$Clicked.on.Ad)

## [1] 0.5
```

The medians for each column is as shown above

d) Range

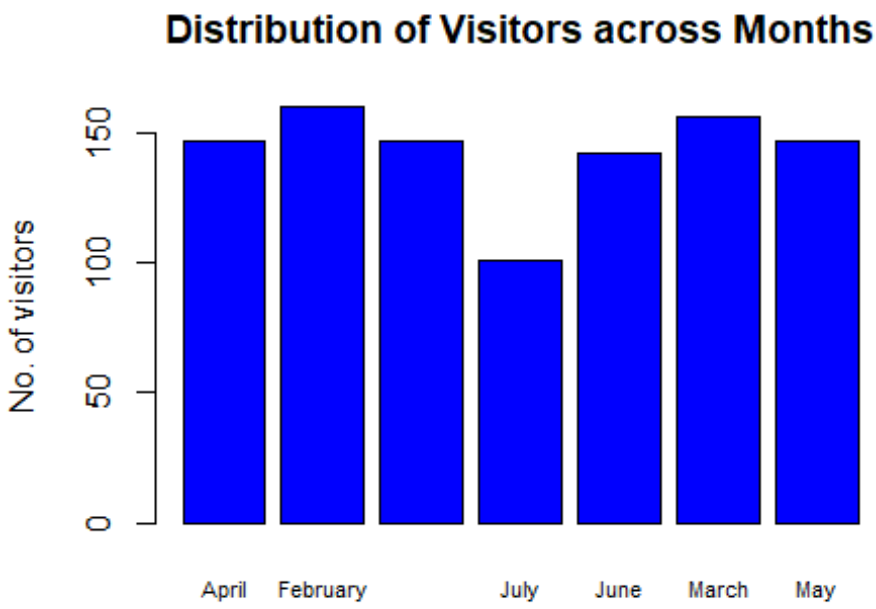
```
# Calculating the ranges
range(data$Daily.Time.Spent.on.Site)

## [1] 32.60 91.43

range(data$Age)

## [1] 19 61

range(data$Area.Income)

## [1] 13996.5 79484.8

range(data$Daily.Internet.Usage)

## [1] 104.78 269.96

range(data$Gender)

## [1] 0 1

range(data$Clicked.on.Ad)
```

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

The codes above show the ranges for each of the columns in our dataset

Plots:

```
# A bar plot showing the distribution of visitors accross different months
barplot(table(data$Month), main = "Distribution of Visitors across Months", y
lab = "No. of visitors", col ="blue", cex.names = 0.7)
```

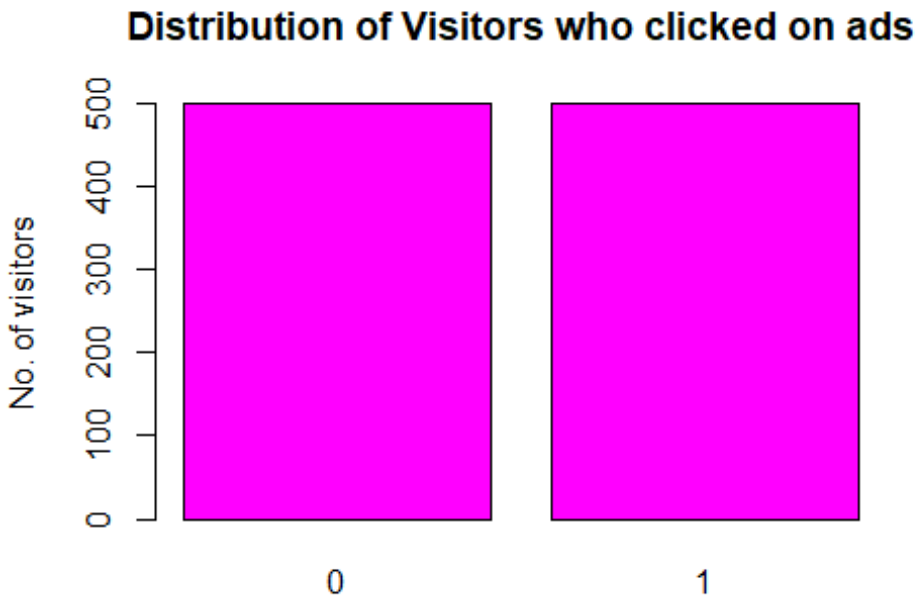


From our plot, we can see that February had the highest site traffic, followed closely by March, then April. July had the least amount of traffic to the site.

```
# A bar plot showing the gender distribution
barplot(table(data$Gender),  col =  c("red" , "green"), ylab = "No. of visito
rs", main = "Distribution of Gender")
```

From the plot, we can see that 0, ie females, caused the highest traffic to the site.

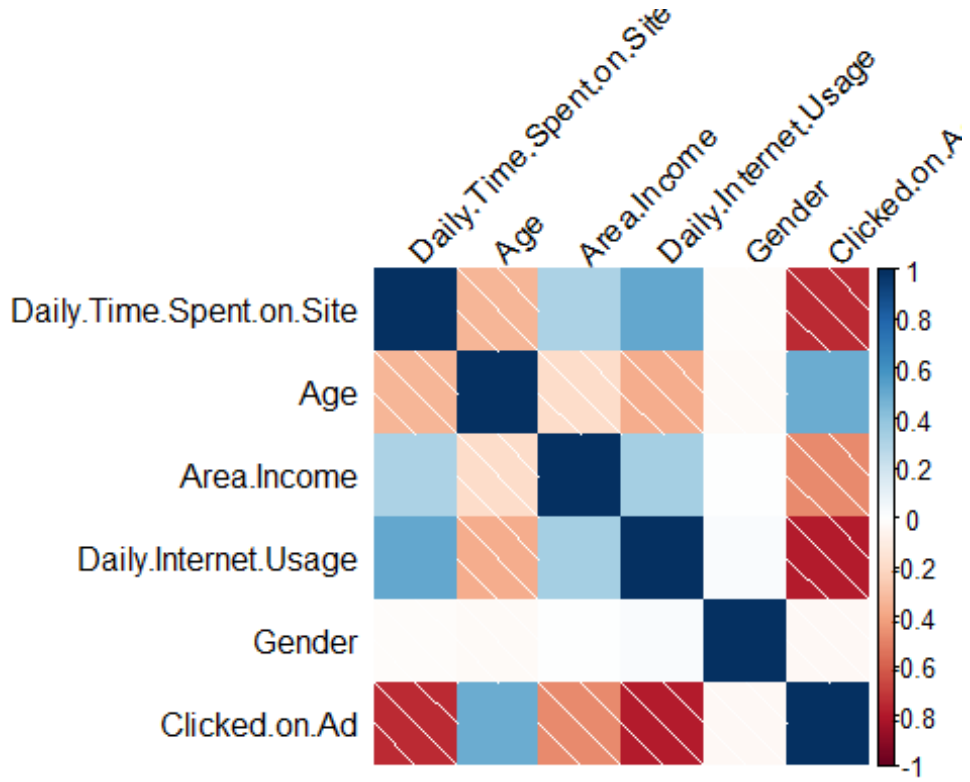
```
# A bar plot showing the distribution of visitors who clicked on the ads
barplot(table(data$Clicked.on.Ad), main = "Distribution of Visitors who click
ed on ads", ylab = "No. of visitors", col ="magenta")
```



From the plot, it seems like there was an equal amount of visitors who clicked on ads as well as those who didn't.

2. Bivariate Analysis

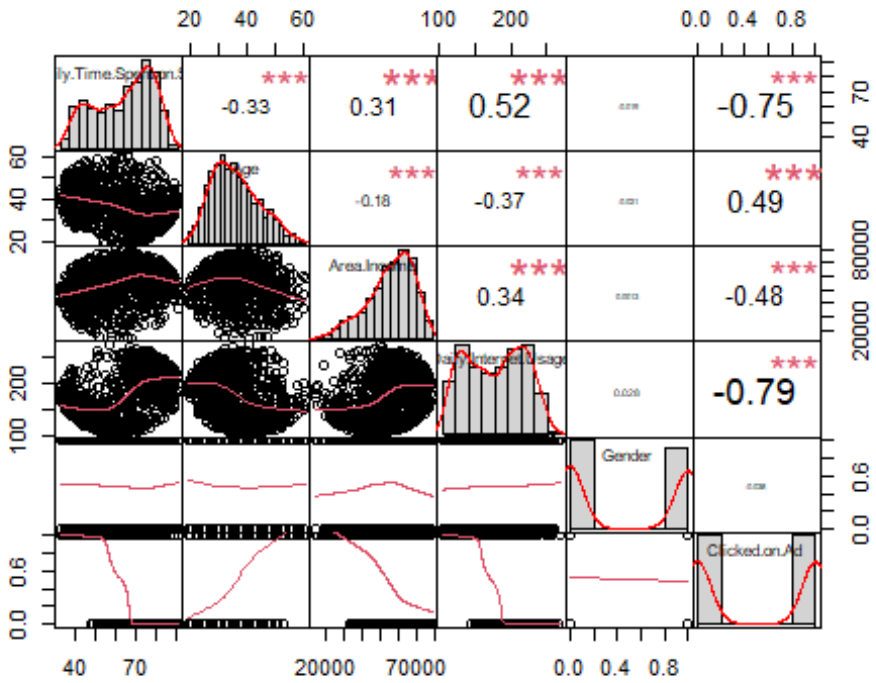
```
#Calculating the correlation between columns

correlation = cor(data_num)

# Creating a correlogram to plot our correlation for better presentation
corrplot(correlation, method="shade", tl.col="black", tl.srt=45)
```



From the correlogram above & using the legend on the right, we can see that:

- Daily time spent on the site has a high negative relationship with whether one clicked on an ad. Thus if one spends alot of time on the site, there is a high chance of them not clicking on an ad.
- There's a low negative relationship between one's age and the daily time spent on the site as well as their daily internet usage. Thus the higher one's age, the less likely they are to spend more time on the site or to have a high internet usage.
- There is a medium negative relationship between one's income and whether they clicked on an ad. Thus, the higher one's income, the less likely they are to click on ad.
- There's a medium positive relationship between one's daily internet usage and the daily time spent on the site. This shows that there's a medium chance that people with a high internet usage would be those spending a lot of time on the site.
- There is also a medium positive relationship between one's age and whether they clicked on ads. Thus the older one's age, the more like they are to click on an ad, however the realtionship is not too strong.

```
# Plotting scatterplots to get the distributions of the columns as well
as the significance value

chart.Correlation(data_num, histogram = TRUE,)
```



From the scatterplot above, we can see the significance levels for the different variables, as well as the scatter plots with the fitted lines. We can also see the different distributions for our datasets.

Conclusion

The most common amount of time spent on the site is 62.26, while the most popular age is 31years. Most of the site visits have an income of 61,833.9.

The city with most visitors is Lisamouth while Czech Republic had the most site visitors.

Most of the visitors were female, while most didn't click on the ads.

The site had most traffic in February.

The internet users that are most likely to click on our client's ads are those who spend very little time online.

Also the lower one's income is, the higher the chances of them clicking on the ads. The older one is, the more like they are to click on an ad, however the relationship is not too strong.

Recommendations

My recommendation as a data scientist, would be for her to do curate the ads to be mostly on courses that may interest the people most likely to click on the ads as outlined above. She can alsoo include other content so as to pique other users to click more on the ads.