Advertising Analysis

Tabitha Kariuki

2022-07-15

# R Programming Exploratory Data Analysis

## 1. Defining the Question

**a) Specifying the Data Analytic Question**

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.

**b) Defining the Metric for Success**

1.Exhaustively apply the exploratory data analysis approaches while defining the question, the metric for success, the context, experimental design taken and the appropriateness of the available data to answer the given question.

2.Perform univariate analysis by calculating and interpreting measures of central tendency for the set of data.

3.Exhaustively perform bivariate analysis by creating relevant visualizations

**c) Understanding the context**

Perform Exploratory Data Analysis for the give data set <http://bit.ly/IPAdvertisingData>

**d) Experimental design taken**

1.Reading and checking our data

2.Clean data by finding and dealing with outliers, anomalies, and missing data within the dataset.

3.Perform univariate and bivariate analysis.

4.From your insights provide a conclusion and recommendation.

**e) Appropriateness of the available data**

The dataset has appropriate columns and rows to answer the questions. The data is relevant for our analysis.

## 2. Importing Libraries

# install.package("data.table") # install package data.table to work with data tables  
library(data.table) # load package  
# install.package("tidyverse") # install packages to work with data frame - extends into visualization  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::between() masks data.table::between()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::first() masks data.table::first()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::last() masks data.table::last()  
## ✖ purrr::transpose() masks data.table::transpose()

## 3. Loading our dataset

# Loading our dataset into our environment  
  
ad <- fread('http://bit.ly/IPAdvertisingData')

## 4. Reading our dataset

# Checking our top rows  
  
head(ad)

## 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 our bottom rows  
  
tail(ad)

## Daily Time Spent on Site Age Area Income Daily Internet Usage  
## 1: 43.70 28 63126.96 173.01  
## 2: 72.97 30 71384.57 208.58  
## 3: 51.30 45 67782.17 134.42  
## 4: 51.63 51 42415.72 120.37  
## 5: 55.55 19 41920.79 187.95  
## 6: 45.01 26 29875.80 178.35  
## Ad Topic Line City Male  
## 1: Front-line bifurcated ability Nicholasland 0  
## 2: Fundamental modular algorithm Duffystad 1  
## 3: Grass-roots cohesive monitoring New Darlene 1  
## 4: Expanded intangible solution South Jessica 1  
## 5: Proactive bandwidth-monitored policy West Steven 0  
## 6: Virtual 5thgeneration emulation Ronniemouth 0  
## Country Timestamp Clicked on Ad  
## 1: Mayotte 2016-04-04 03:57:48 1  
## 2: Lebanon 2016-02-11 21:49:00 1  
## 3: Bosnia and Herzegovina 2016-04-22 02:07:01 1  
## 4: Mongolia 2016-02-01 17:24:57 1  
## 5: Guatemala 2016-03-24 02:35:54 0  
## 6: Brazil 2016-06-03 21:43:21 1

# Checking the shape of our data  
  
dim(ad)

## [1] 1000 10

We have 1000 rows and 10 columns

# Checking the class/datatypes  
  
str(ad)

## 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>

# checking the attributes of our dataset  
  
class(ad)

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

## 5. Data Cleaning

# Sum of null values in each column using the function colSums()  
  
colSums(is.na(ad))

## 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

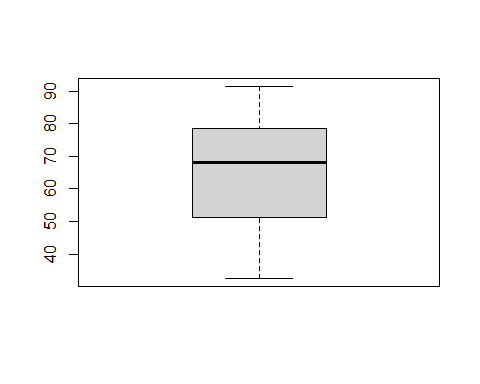
**There are no missing values in our dataset**

# Now lets find the duplicated rows in the dataset   
# and assign to a variable duplicated\_rows below  
  
duplicated\_rows <- ad[duplicated(ad),]  
  
# Lets print out the variable duplicated\_rows and see these duplicated rows   
  
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...

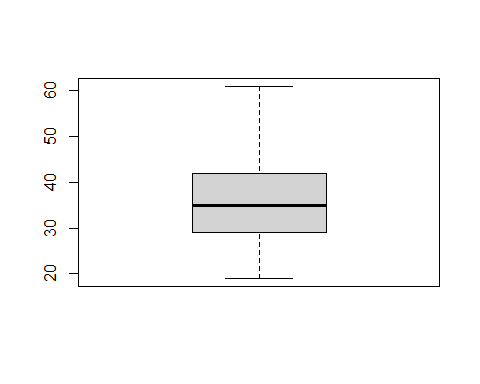
**There are no duplicated rows**

# Checking for outliers in the Daily Time Spent on Site column  
  
boxplot(ad$'Daily Time Spent on Site')



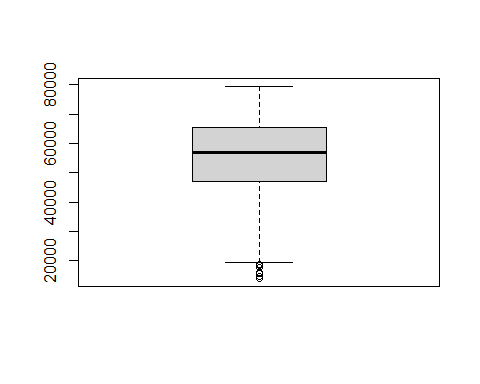
**There are no outliers in the ‘Daily Time Spent on Site’ column**

# Checking for outliers in the age column  
  
boxplot(ad$'Age')



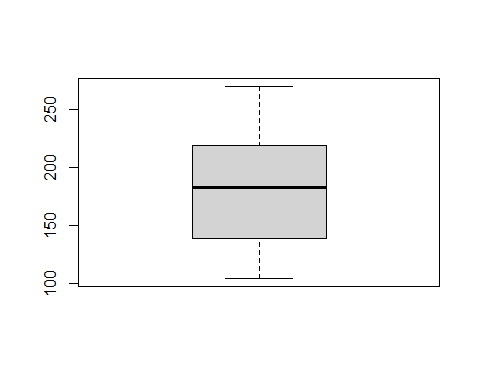
**There are no outliers in the age column**

# Checking for outliers in the Area Income column  
  
boxplot(ad$'Area Income')



**There are outliers in the ‘area income’ column. However we will not be dropping them since it is true representation of individual’s income**

# Checking for outliers in the Daily Internet Usage column  
  
boxplot(ad$'Daily Internet Usage')



**There are no outliers in the ‘Daily Internet usage’ column**

# Exploratory Data Analysis

## 6. Univariate Analysis

# Summary statistics of our data  
  
summary(ad)

## 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 02:52:10.00 Min. :0.0   
## 1st Qu.:2016-02-18 02:55:42.00 1st Qu.:0.0   
## Median :2016-04-07 17:27:29.50 Median :0.5   
## Mean :2016-04-10 10:34:06.64 Mean :0.5   
## 3rd Qu.:2016-05-31 03:18:14.00 3rd Qu.:1.0   
## Max. :2016-07-24 00:22:16.00 Max. :1.0

**Mean**

mean(ad$"Daily Time Spent on Site")

## [1] 65.0002

**The average time spent on the site is 65 minutes.**

mean(ad$"Age")

## [1] 36.009

**The average age of repondents is 36 years.**

mean(ad$"Area Income")

## [1] 55000

**The average income of repondents is 55000**

mean(ad$"Daily Internet Usage")

## [1] 180.0001

**The average internet usage is 180.0 units**

**Mode**

# Unfotunately, R does not have a standard in-built function to calculate mode so we have to build one  
# We create the mode function that will perform our mode operation for us  
  
getmode <- function(v) {  
 uniqv <- unique(v)  
 uniqv[which.max(tabulate(match(v, uniqv)))]}

getmode(ad$Age)

## [1] 31

**Most frequent age is 31 years**

getmode(ad$`Daily Time Spent on Site`)

## [1] 62.26

**Most frequent daily time spent is 62.26 minutes**

getmode(ad$`Area Income`)

## [1] 61833.9

**Most common area income is 61833.9**

getmode(ad$`Daily Internet Usage`)

## [1] 167.22

**Most frequent units used for daily internet usage is 167.22.**

**Median** **The median is the middle number in a sorted, ascending or descending list of numbers**

median(ad$`Daily Time Spent on Site`)

## [1] 68.215

median(ad$Age)

## [1] 35

median(ad$`Area Income`)

## [1] 57012.3

median(ad$`Daily Internet Usage`)

## [1] 183.13

**Min and Max Values/Otherwise known as Range**

**Showing the highest and the least values in our numerical data**

range(ad$Age)

## [1] 19 61

range(ad$`Daily Time Spent on Site`)

## [1] 32.60 91.43

range(ad$`Area Income`)

## [1] 13996.5 79484.8

range(ad$`Daily Internet Usage`)

## [1] 104.78 269.96

**Quantiles**

Getting the first and the third quantile together with the range and the median using the quantile() function

quantile(ad$Age)

## 0% 25% 50% 75% 100%   
## 19 29 35 42 61

quantile(ad$`Daily Time Spent on Site`)

## 0% 25% 50% 75% 100%   
## 32.6000 51.3600 68.2150 78.5475 91.4300

quantile(ad$`Area Income`)

## 0% 25% 50% 75% 100%   
## 13996.50 47031.80 57012.30 65470.64 79484.80

quantile(ad$`Daily Internet Usage`)

## 0% 25% 50% 75% 100%   
## 104.7800 138.8300 183.1300 218.7925 269.9600

**Standard Deviation**

A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.

sd(ad$`Daily Time Spent on Site`)

## [1] 15.85361

sd(ad$Age)

## [1] 8.785562

sd(ad$`Area Income`)

## [1] 13414.63

sd(ad$`Daily Internet Usage`)

## [1] 43.90234

**Variance**

The variance is a numerical measure of how the data values is dispersed around the mean.

var(ad$`Daily Time Spent on Site`)

## [1] 251.3371

var(ad$Age)

## [1] 77.18611

var(ad$`Area Income`)

## [1] 179952406

var(ad$`Daily Internet Usage`)

## [1] 1927.415

**Frequencies**

# Gender Frequency Table  
# 0 for not male while 1 is male  
  
gender = table(ad$Male)  
gender

##   
## 0 1   
## 519 481

**519 respondents are not Male while 481 are male**

# city Frequency Table  
  
city = table(ad$City)  
  
# Arranging cities from the most frequent and displaying the first 6 rows   
  
highestcity <- sort(city, decreasing = TRUE)  
head(highestcity)

##   
## Lisamouth Williamsport Benjaminchester East John East Timothy   
## 3 3 2 2 2   
## Johnstad   
## 2

# country Frequency Table  
  
country = table(ad$Country)  
  
# Arranging countries from the least frequent and displaying the first 6 rows   
  
countries <- sort(country, increasing = TRUE)  
head(countries)

##   
## Aruba   
## 1   
## Bermuda   
## 1   
## British Indian Ocean Territory (Chagos Archipelago)   
## 1   
## Cape Verde   
## 1   
## Germany   
## 1   
## Jordan   
## 1

# clicked on ad Frequency Table  
  
clickad = table(ad$`Clicked on Ad`)  
clickad

##   
## 0 1   
## 500 500

# clicked on ad Frequency Table  
  
clickad = table(ad$`Clicked on Ad`)  
clickad

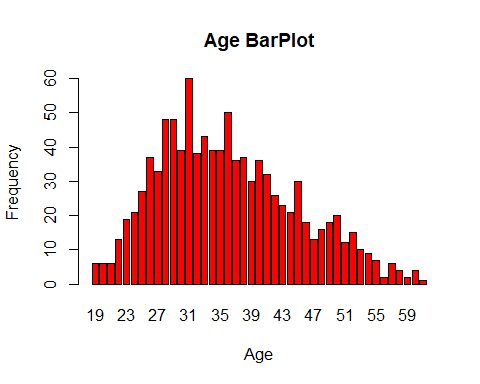
##   
## 0 1   
## 500 500

**Half of the respondents clicked on ad**

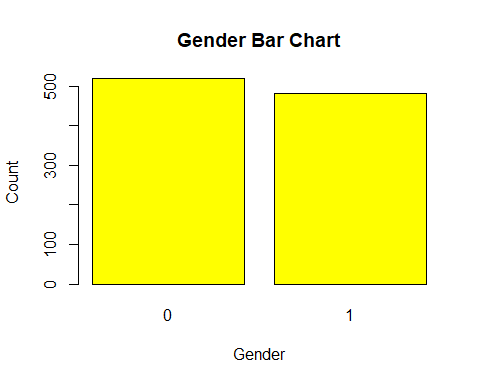
**Barplots**

The box plot of an observation variable is a graphical representation based on its quantiles, as well as its smallest and largest values. It attempts to provide a visual shape of the data distribution.

# Fits we get the frequency distribution table  
  
age <- table(ad$Age)  
  
# Then we plot a bar chart   
  
barplot(age, xlab ='Age', ylab ='Frequency', main ='Age BarPlot', col = 'red')

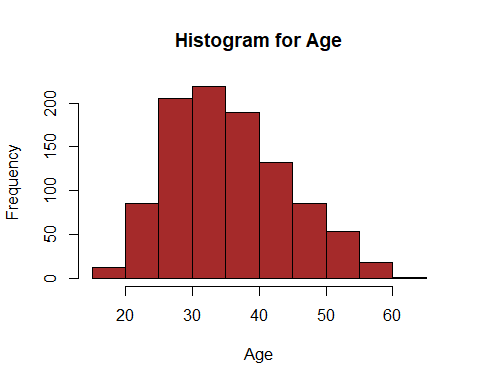


# Gender barplot  
  
barplot(gender, xlab = 'Gender', ylab = 'Count', main = 'Gender Bar Chart', col = "yellow")



**Histograms**

# Plot a histogram for the age column  
  
hist(ad$Age, xlab = 'Age', main = 'Histogram for Age', col = 'brown')



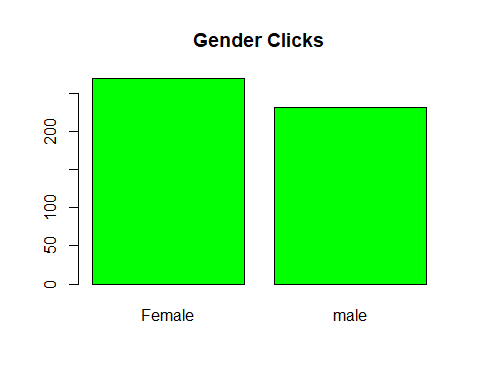
## 7. BiVariate Analysis\*\*

**Which Gender clicked the most ads**

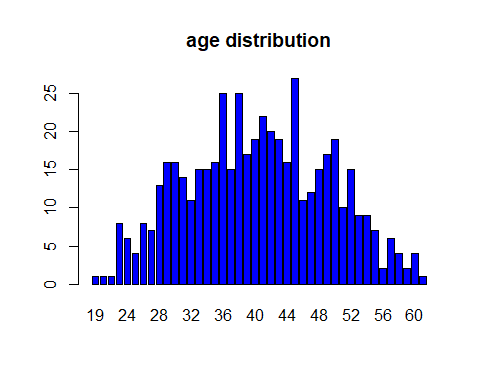
# Creating a dataframe for those who clicked the ad  
  
clicked <- ad[ad$'Clicked on Ad'==1,]  
head(clicked)

## Daily Time Spent on Site Age Area Income Daily Internet Usage  
## 1: 66.00 48 24593.33 131.76  
## 2: 47.64 49 45632.51 122.02  
## 3: 69.57 48 51636.92 113.12  
## 4: 42.95 33 30976.00 143.56  
## 5: 63.45 23 52182.23 140.64  
## 6: 55.39 37 23936.86 129.41  
## Ad Topic Line City Male  
## 1: Reactive local challenge Port Jefferybury 1  
## 2: Centralized neutral neural-net West Brandonton 0  
## 3: Centralized content-based focus group West Katiefurt 1  
## 4: Grass-roots coherent extranet West William 0  
## 5: Persistent demand-driven interface New Travistown 1  
## 6: Customizable multi-tasking website West Dylanberg 0  
## Country Timestamp Clicked on Ad  
## 1: Australia 2016-03-07 01:40:15 1  
## 2: Qatar 2016-03-16 20:19:01 1  
## 3: Egypt 2016-06-03 01:14:41 1  
## 4: Barbados 2016-03-24 09:31:49 1  
## 5: Spain 2016-03-09 03:41:30 1  
## 6: Palestinian Territory 2016-01-30 19:20:41 1

genderclicks<- table(clicked$Male)  
label<- c("Female","male")  
barplot(genderclicks,names.arg=label,main="Gender Clicks", col = 'green')



ageDist<- table(clicked$Age)  
barplot(ageDist,main="age distribution", col = 'blue')



**Covariance**

# We can find the covariance between age and the daily time spent on the site  
  
age <- ad$Age  
time <- ad$"Daily Time Spent on Site"  
  
cov(age, time)

## [1] -46.17415

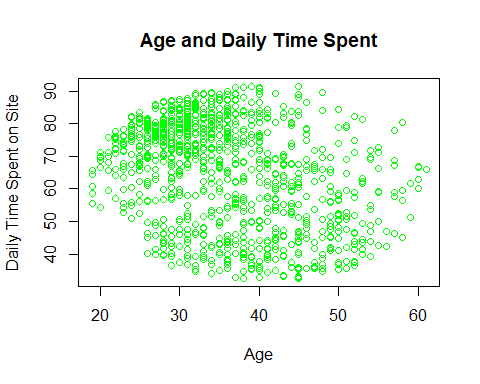
**There is a negative covariance between age and the daily time spent on the site which means that the older a person is, the less time they spend on the site daily.**

# We can find the covariance between age and the internet units  
  
age <- ad$Age  
units <- ad$`Daily Internet Usage`  
  
cov(age, units)

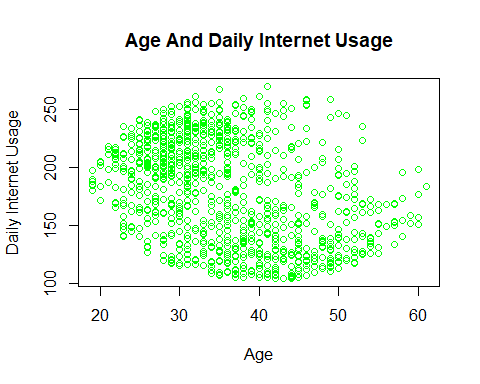
## [1] -141.6348

**There is a negative covariance between age and the daily internet usage on the site which means that the older a person gets, the less time they spend on daily internet usage.**

# Scatter plot showing distribution of age and time spent on site  
  
plot(age, time, xlab = 'Age', ylab = 'Daily Time Spent on Site', main = 'Age and Daily Time Spent', col = 'green')



# Scatter plot showing distribution of age and Internet usage  
  
plot(age, units, xlab = 'Age', ylab = 'Daily Internet Usage', main = 'Age And Daily Internet Usage', col = "green")



## 8. Conclusion

1. There were more females than males in our data.
2. The data had 500 individuals who clicked on the ads while 500 individuals did not click on the ads.
3. Czech Republic and France both had the highest number of respondents.
4. The average area income is 55000.
5. The average age of most audience is 36 years with most of the audience being around the age of 31.
6. Lisamouth and Williamsport cities both had the highest number of individuals (3) in the dataset.
7. There are more females visiting the site compared to males as well as clicking the ads.

## 9. Recommendations

1. Individuals who are of the female gender and are between 28 and 36 years old were the most in our data set, therefore she should creates an ad that targets these individuals
2. Most of the those who click on the ad have an area income of 55000, so maybe reevaluate the prices to accommodate other income levels.