# Ad Optimization in R

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.

## Research Question

A model that assist the enterpreneur best decide which factors to use in rolling out their advertisements to get highest engagement on them(calculated by number of users that click on the ads) ## Metric of Success

**Loading dataset**

ad = read.csv("http://bit.ly/IPAdvertisingData")

**Previewing the Dataset**

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

tail(ad)

## 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

**Data Cleaning**

#Checking for null values in our dataset  
  
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

our dataset does not contain any missing values.

# Checking for duplicates   
  
anyDuplicated(ad)

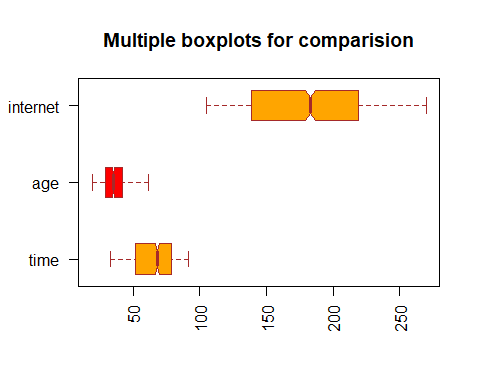
## [1] 0

Our dataset does not contain any duplicates.

str(ad)

## '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 : Factor w/ 1000 levels "Adaptive 24hour Graphic Interface",..: 92 465 567 904 767 806 223 724 108 455 ...  
## $ City : Factor w/ 969 levels "Adamsbury","Adamside",..: 962 904 112 940 806 283 47 672 885 713 ...  
## $ Male : int 0 1 0 1 0 1 0 1 1 1 ...  
## $ Country : Factor w/ 237 levels "Afghanistan",..: 216 148 185 104 97 159 146 13 83 79 ...  
## $ Timestamp : Factor w/ 1000 levels "2016-01-01 02:52:10",..: 440 475 368 57 768 690 131 334 549 942 ...  
## $ Clicked.on.Ad : int 0 0 0 0 0 0 0 1 0 0 ...

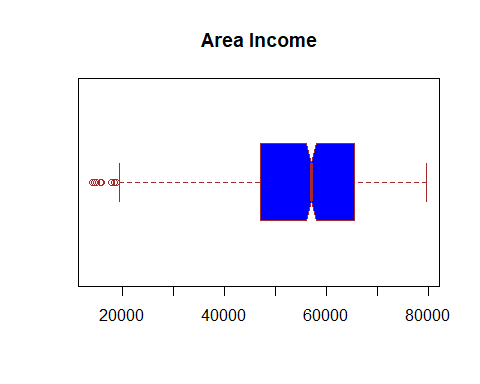
# Checking for outliers  
  
# prepare the data  
time <- ad$Daily.Time.Spent.on.Site  
age <- ad$Age  
internet <- ad$Daily.Internet.Usage  
  
boxplot(time, age, internet,  
main = "Multiple boxplots for comparision",  
at = c(1,3,5),  
names = c("time", "age", "internet"),  
las = 2,  
col = c("orange","red"),  
border = "brown",  
horizontal = TRUE,  
notch = TRUE  
)



The “Daily time spent on Site”, “Age” and “Daily Internet Usage” columns do not contain any outliers.

The values on the income column were far too different to be plotted with the other columns while still making the box plots make sense. A box plot for income values is done below.

boxplot(ad$Area.Income,  
main = "Area Income",  
col = "blue",  
border = "brown",  
horizontal = TRUE,  
notch = TRUE  
)

 The “Area Income” has several outliers. However, since these represent income earned by different people in different geographical areas, the outliers are assumed to be factual and will not be removed or replaced.

# Exploratory Data Analysis

## Univariate Exploratory Data Analysis

Calculating/identifying several measures of central tendency.

To start of we will create a function for calculating mode.

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

**Finding the mode**

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

## [1] 62.26

getmode(ad$Age)

## [1] 31

getmode(ad$Area.Income)

## [1] 61833.9

getmode(ad$Daily.Internet.Usage)

## [1] 167.22

The mode for daily time spent on the site is 62.26 The mode for age is 31 The mode for area income is 6.1833910^{4} The mode for daily internet usage is 167.22

**Finding the mean**

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

## [1] 65.0002

mean(ad$Age)

## [1] 36.009

mean(ad$Area.Income)

## [1] 55000

mean(ad$Daily.Internet.Usage)

## [1] 180.0001

The mean daily time spent on the site is 65.0002 The mean for the age is 36.009 The mean area income is 5.510^{4} The mean daily internet usage is 180.0001

**Finding the Median**

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

The median daily time spent on the site is 68.215 The median for the age is 35 The median area income is 5.7012310^{4} The median daily internet usage is 183.13

**Calculating Measures of Dispersion**

**Finding the maximum values**

max(ad$Daily.Time.Spent.on.Site)

## [1] 91.43

max(ad$Age)

## [1] 61

max(ad$Area.Income)

## [1] 79484.8

max(ad$Daily.Internet.Usage)

## [1] 269.96

The maximum daily time spent on the site is 91.43 The maximum for the age is 61 The maximum area income is 7.9484810^{4} The maximum daily internet usage is 269.96

**Finding the minimum values**

min(ad$Daily.Time.Spent.on.Site)

## [1] 32.6

min(ad$Age)

## [1] 19

min(ad$Area.Income)

## [1] 13996.5

min(ad$Daily.Internet.Usage)

## [1] 104.78

The minmum daily time spent on the site is 32.6 The minmum for the age is 19 The minmum area income is 1.3996510^{4} The minmum daily internet usage is 104.78

**Finding the Range**

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

## [1] 32.60 91.43

range(ad$Age)

## [1] 19 61

range(ad$Area.Income)

## [1] 13996.5 79484.8

range(ad$Daily.Internet.Usage)

## [1] 104.78 269.96

The range daily time spent on the site is 32.6, 91.43 The range for the age is 19, 61 The range area income is 1.3996510^{4}, 7.9484810^{4} The range daily internet usage is 104.78, 269.96

**Finding the Variance**

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

The variance for daily time spent on the site is 251.3370949 The variance for the age is 77.1861051 The variance for area income is 1.799524110^{8} The variance for daily internet usage is 1927.4153962

**Finding the Standard Deviation**

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

The Standard deviation for daily time spent on the site is 15.8536146 The Standard deviation for the age is 8.7855623 The Standard deviation for area income is 1.341463410^{4} The Standard deviation for daily internet usage is 43.9023393

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

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

quantile(ad$Age)

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

quantile(ad$Area.Income)

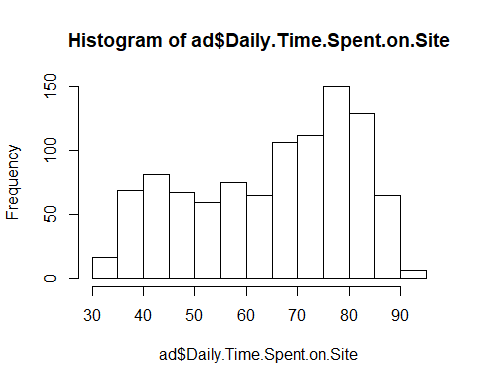
## 0% 25% 50% 75% 100%   
## 13996.50 47031.80 57012.30 65470.64 79484.80

quantile(ad$Daily.Internet.Usage)

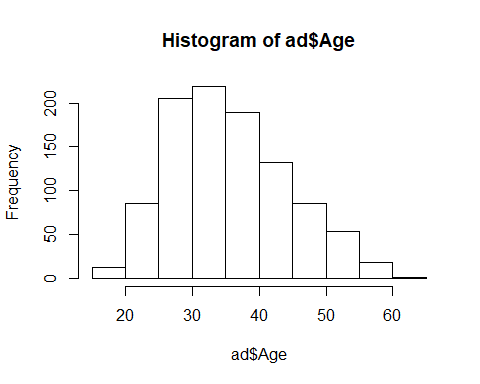
## 0% 25% 50% 75% 100%   
## 104.7800 138.8300 183.1300 218.7925 269.9600

**Visualisations**

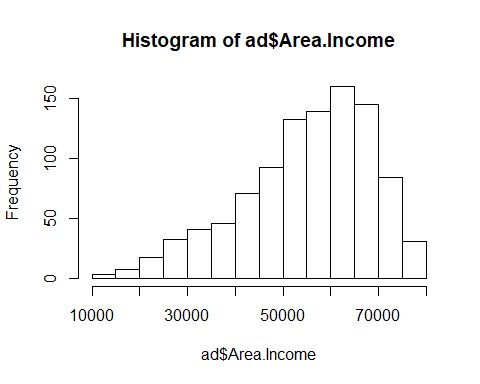
hist(ad$Daily.Time.Spent.on.Site)



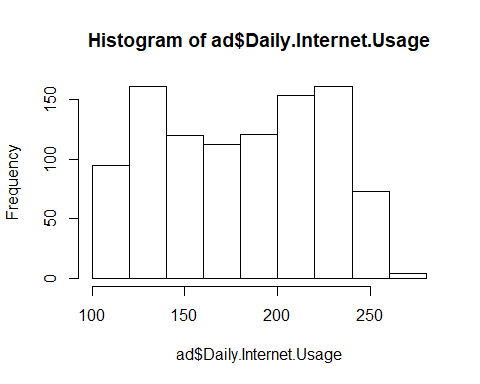
hist(ad$Age)



hist(ad$Area.Income)



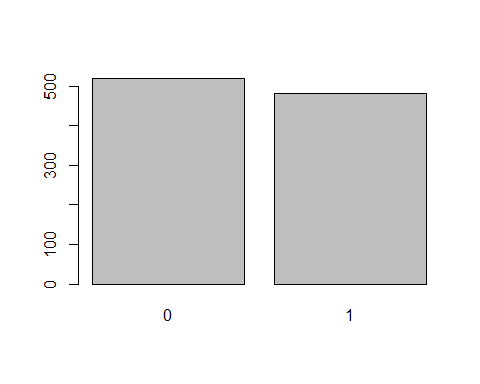
hist(ad$Daily.Internet.Usage)



table(ad$Male)

##   
## 0 1   
## 519 481

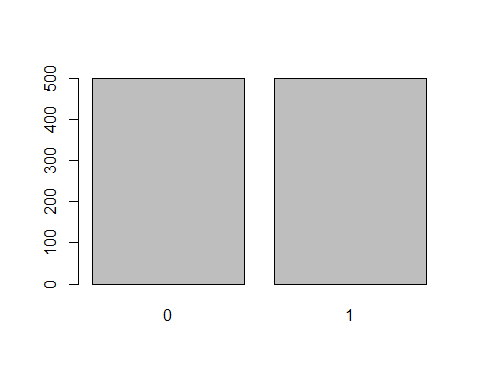
barplot(table(ad$Male))

 The dataset has 519 females and 481 males.

table(ad$Clicked.on.Ad)

##   
## 0 1   
## 500 500

barplot(table(ad$Clicked.on.Ad))

 500 people clicked on the ad, while 500 others did not click on the ad.

# Bivariate Analysis

# Creating variables to used in checking for covariance  
time <- ad$Daily.Time.Spent.on.Site  
age <- ad$Age  
income <- ad$Area.Income  
usage <- ad$Daily.Internet.Usage  
gender <- ad$Male  
click <- ad$Clicked.on.Ad

**Covariance between likelihood of clicking the ad and other variables**

cov(click,time)

## [1] -5.933143

cov(click,age)

## [1] 2.164665

cov(click,income)

## [1] -3195.989

cov(click,usage)

## [1] -17.27409

cov(click,gender)

## [1] -0.00950951

The chance of clicking an ad has a positive linear relationship with the age of the site user. **Correlation between likelihood of clicking the ad and other variables**

cor(click,time)

## [1] -0.7481166

cor(click,age)

## [1] 0.4925313

cor(click,income)

## [1] -0.4762546

cor(click,usage)

## [1] -0.7865392

cor(click,gender)

## [1] -0.03802747

There is a high negative correlation between chance of clicking the ad with time spent on the site and daily internet usage There is a week correlation between chance of clicking on an ad and the age of the user 0.4925313

my\_data <- ad[, c(1,2,3,4,7,10)]  
  
head(my\_data)

## 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

res <- cor(my\_data)  
round(res, 2)

## Daily.Time.Spent.on.Site Age Area.Income  
## Daily.Time.Spent.on.Site 1.00 -0.33 0.31  
## Age -0.33 1.00 -0.18  
## Area.Income 0.31 -0.18 1.00  
## Daily.Internet.Usage 0.52 -0.37 0.34  
## Male -0.02 -0.02 0.00  
## Clicked.on.Ad -0.75 0.49 -0.48  
## Daily.Internet.Usage Male Clicked.on.Ad  
## Daily.Time.Spent.on.Site 0.52 -0.02 -0.75  
## Age -0.37 -0.02 0.49  
## Area.Income 0.34 0.00 -0.48  
## Daily.Internet.Usage 1.00 0.03 -0.79  
## Male 0.03 1.00 -0.04  
## Clicked.on.Ad -0.79 -0.04 1.00

cor(my\_data, use = "complete.obs")

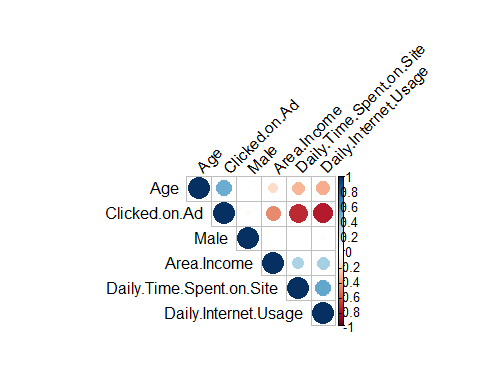
## Daily.Time.Spent.on.Site Age Area.Income  
## Daily.Time.Spent.on.Site 1.00000000 -0.33151334 0.310954413  
## Age -0.33151334 1.00000000 -0.182604955  
## Area.Income 0.31095441 -0.18260496 1.000000000  
## Daily.Internet.Usage 0.51865848 -0.36720856 0.337495533  
## Male -0.01895085 -0.02104406 0.001322359  
## Clicked.on.Ad -0.74811656 0.49253127 -0.476254628  
## Daily.Internet.Usage Male Clicked.on.Ad  
## Daily.Time.Spent.on.Site 0.51865848 -0.018950855 -0.74811656  
## Age -0.36720856 -0.021044064 0.49253127  
## Area.Income 0.33749553 0.001322359 -0.47625463  
## Daily.Internet.Usage 1.00000000 0.028012326 -0.78653918  
## Male 0.02801233 1.000000000 -0.03802747  
## Clicked.on.Ad -0.78653918 -0.038027466 1.00000000

library(corrplot)

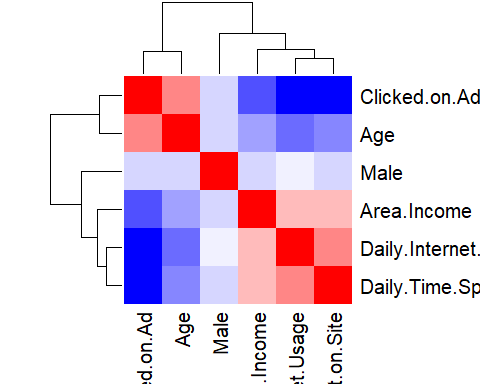
## Warning: package 'corrplot' was built under R version 3.6.3

## corrplot 0.84 loaded

corrplot(res, type = "upper", order = "hclust",   
 tl.col = "black", tl.srt = 45)



# Get some colors  
col<- colorRampPalette(c("blue", "white", "red"))(20)  
heatmap(x = res, col = col, symm = TRUE)



# Summary

From the above analysis, it was observed that given all the factors listed above, there was an equal chance of the users clicking on the ad or choosing not to click on the ad.

The chance of clicking on the ads was also highly negatively correlated with the following factors; daily internet usage and daily internet usage.