velvin-week12-ip

Velsamkul

4/6/2021

# Research Question

A Kenyan entrepreneur has created an online cryptography course and would wantto 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.

## 1.a) Specifying the Question

To carry out data cleaning and exploratory data analysis To identify which individuals are most likely to click on the ads of a kenyan entrepreneur based on the given data and come up with recommendations.

##b) Defining the metric for success

Success will have been achieved when I complete exploratory data analysis, have clear plots and finaly identify individuals who are likely to click on ads.

##c) Recording the Experimental Design

1. Loading the data
2. Checking the data
3. Tidying the data
4. Univariate Analysis
5. Bivariate Analysis
6. Challenging the solution
7. Recommendations
8. Follow up questions

##4 Checking Data Relevance

The data provided was found to be relevant for our analysis because it contains values that are useful.

## 2.Loading and reading Our Datasets

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.1.0 v dplyr 1.0.5  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

ads <- read.csv("C:/Users/Admin/Downloads/advertising.csv")  
view(ads)

Here is a summary of the data

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 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  
## Length:1000 Min. :0.0   
## Class :character 1st Qu.:0.0   
## Mode :character Median :0.5   
## Mean :0.5   
## 3rd Qu.:1.0   
## Max. :1.0

Preview of the the first and last few rows.

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

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

Shape of the data.

cat("Rows in dataset:", nrow(ads), "\nCols in dataset:", ncol(ads))

## Rows in dataset: 1000   
## Cols in dataset: 10

cat("\nThe dimension of the dataset is:", dim(ads))

##   
## The dimension of the dataset is: 1000 10

Structure of our dataframe

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 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 : 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 ...

As seen, there are 1000 records and 10 variables. 3 variables of the numeric, 3 integer types, 4 character types including the date and time which will be converted to the standard format.

#Data cleaning

*Converting the date and time*

class(ads$Timestamp)

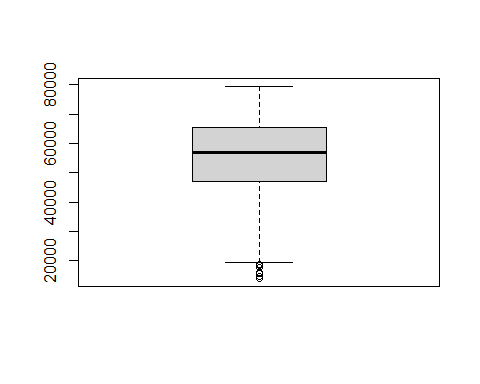
## [1] "character"

ads$Timestamp <- strptime(paste( ads$Timestamp), format = "%Y-%m-%d %H:%M:%S",tz="UTC")   
class(ads$Timestamp)

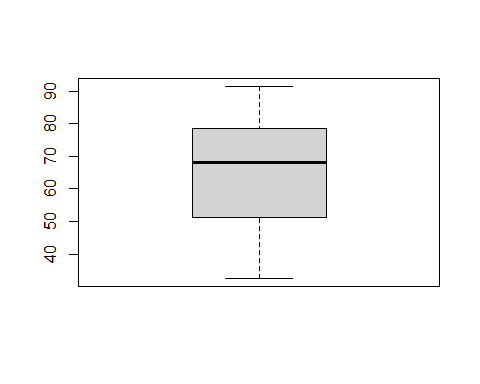
## [1] "POSIXlt" "POSIXt"

*Checking for outliers on the numerical columns using boxplots*

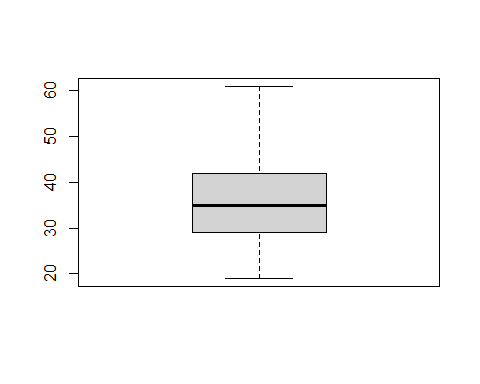
boxplot(ads$Area.Income)



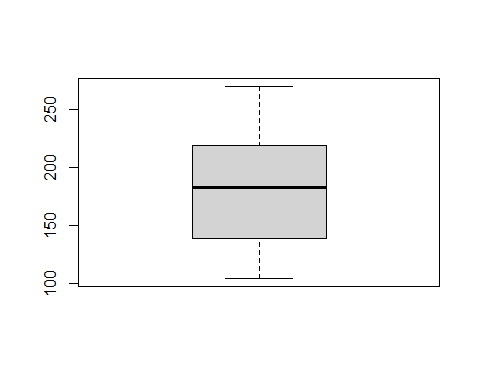
boxplot(ads$Daily.Time.Spent.on.Site)



boxplot(ads$Age)

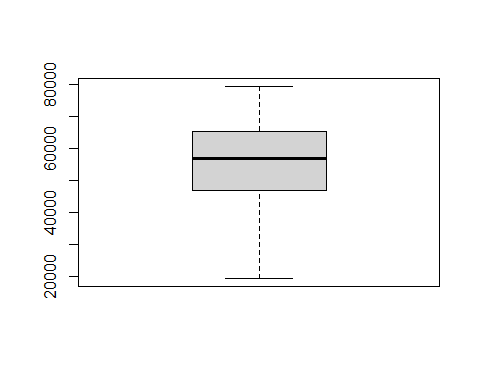


boxplot(ads$Daily.Internet.Usage)

 Only Area had outliers, Area.

*Dealing with outliers;*

bench <- 47032 - 1.5 \* IQR(ads$Area.Income)   
ads$Area.Income[ads$Area.Income < bench]<- bench  
  
boxplot(ads$Area.Income)

 The boxplot shows that the outliers in that column have been removed.

*Checking for duplicates*

sum(duplicated(ads))

## [1] 0

There are no duplicates in our data

*Checking for missing values*

colSums(is.na(ads))

## 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

The dataset has no missing values in any of the columns.

# Exploratory Data Analysis

install.packages("dataMaid", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/Admin/Documents/R/win-library/4.0'  
## (as 'lib' is unspecified)

## Warning in readRDS(dest): lzma decoding result 10

## Warning: unable to access index for repository http://cran.us.r-project.org/bin/windows/contrib/4.0:  
## cannot open URL 'http://cran.us.r-project.org/bin/windows/contrib/4.0/PACKAGES'

## installing the source package 'dataMaid'

## Warning in download.file(url, destfile, method, mode = "wb", ...):  
## InternetOpenUrl failed: 'The server name or address could not be resolved'

## Error in download.file(url, destfile, method, mode = "wb", ...) :   
## cannot open URL 'http://cran.us.r-project.org/src/contrib/dataMaid\_1.4.0.tar.gz'

## Warning in download.packages(pkgs, destdir = tmpd, available = available, :  
## download of package 'dataMaid' failed

install.packages("inspectdf", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/Admin/Documents/R/win-library/4.0'  
## (as 'lib' is unspecified)

## Warning: unable to access index for repository http://cran.us.r-project.org/bin/windows/contrib/4.0:  
## cannot open URL 'http://cran.us.r-project.org/bin/windows/contrib/4.0/PACKAGES'

## Package which is only available in source form, and may need  
## compilation of C/C++/Fortran: 'inspectdf'

## These will not be installed

*Calling the libraries*

library(dplyr)  
library(inspectdf)

The 2 packages will give us more insights on our data.

inspect\_cat(ads)

## # A tibble: 3 x 5  
## col\_name cnt common common\_pcnt levels   
## <chr> <int> <chr> <dbl> <named list>   
## 1 Ad.Topic.Line 1000 Adaptive 24hour Graphic Int~ 0.1 <tibble [1,000 x~  
## 2 City 969 Lisamouth 0.3 <tibble [969 x 3~  
## 3 Country 237 Czech Republic 0.900 <tibble [237 x 3~

**common\_pcnt**, the percentage of each column occupied by the most common level shown in **common.**

#*Univariate Analysis*

Range of Time Spent on Site by users

site.time.range <- range(ads$Daily.Time.Spent.on.Site)  
cat("The Range of Time Spent on Site by users is:",site.time.range)

## The Range of Time Spent on Site by users is: 32.6 91.43

Range of Daily Internet Usage

internet.time.range <- range(ads$Daily.Internet.Usage)  
cat("The Range of Daily Internet Usage is:", internet.time.range)

## The Range of Daily Internet Usage is: 104.78 269.96

Range of Age

age.range <- range(ads$Age)  
cat("The Range of Users' age is:",age.range)

## The Range of Users' age is: 19 61

Range of Income

income.range <- range(ads$Area.Income)  
cat("The Range of Users' income is:",income.range)

## The Range of Users' income is: 19373.75 79484.8

#Bivariate Analysis visualization

check for correlation between the different columns and the target variable Clicked.On.Ad.

inspect\_cor(ads, df2 = NULL, method = "pearson", with\_col = 'Clicked.on.Ad', alpha = 0.05)

## Warning: Columns with 0 variance found: Male, Clicked.on.Ad

## # A tibble: 5 x 7  
## col\_1 col\_2 corr p\_value lower upper pcnt\_nna  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Clicked.on.Ad Daily.Internet.Usage -0.787 3.74e-136 -0.809 -0.762 100  
## 2 Clicked.on.Ad Daily.Time.Spent.on.~ -0.748 2.29e-123 -0.774 -0.719 100  
## 3 Clicked.on.Ad Age 0.493 1.55e- 54 0.444 0.538 100  
## 4 Clicked.on.Ad Area.Income -0.477 3.04e- 51 -0.523 -0.428 100  
## 5 Clicked.on.Ad Male -0.0380 2.30e- 1 -0.0998 0.0240 100

The summary above shows Pearson’s correlation coefficients for all the numeric columns, compared against the ‘Clicked.On.Ads’ column. Tere are negative correlation values for ‘Daily.Internet.Usage’, ‘Daily.Time.Spent.on.Site’, ‘Area Income’. The only positive correlation is between Clicked.On.Ad’ and ‘Age’. Generally, there is high correlation.

inspect\_cor(ads, df2 = NULL, method = "pearson", alpha = 0.05)

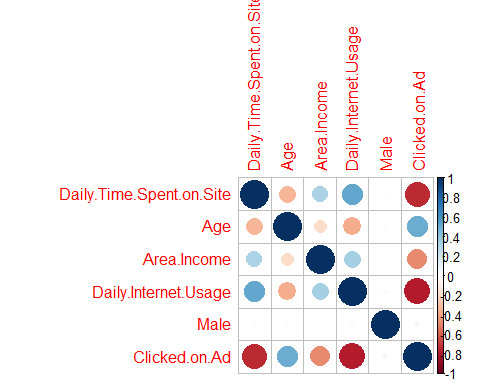
## Warning: Columns with 0 variance found: Male, Clicked.on.Ad

## # A tibble: 15 x 7  
## col\_1 col\_2 corr p\_value lower upper pcnt\_nna  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Clicked.on.Ad Daily.Internet.U~ -0.787 3.74e-136 -0.809 -0.762 100  
## 2 Clicked.on.Ad Daily.Time.Spent~ -0.748 2.29e-123 -0.774 -0.719 100  
## 3 Daily.Internet~ Daily.Time.Spent~ 0.519 2.80e- 60 0.472 0.563 100  
## 4 Clicked.on.Ad Age 0.493 1.55e- 54 0.444 0.538 100  
## 5 Clicked.on.Ad Area.Income -0.477 3.04e- 51 -0.523 -0.428 100  
## 6 Daily.Internet~ Age -0.367 4.38e- 31 -0.420 -0.312 100  
## 7 Daily.Internet~ Area.Income 0.339 8.86e- 27 0.283 0.393 100  
## 8 Age Daily.Time.Spent~ -0.332 1.22e- 25 -0.386 -0.275 100  
## 9 Area.Income Daily.Time.Spent~ 0.312 6.35e- 23 0.255 0.367 100  
## 10 Area.Income Age -0.182 8.44e- 9 -0.242 -0.122 100  
## 11 Clicked.on.Ad Male -0.0380 2.30e- 1 -0.0998 0.0240 100  
## 12 Male Daily.Internet.U~ 0.0280 3.76e- 1 -0.0340 0.0898 100  
## 13 Male Age -0.0210 5.06e- 1 -0.0829 0.0410 100  
## 14 Male Daily.Time.Spent~ -0.0190 5.50e- 1 -0.0808 0.0431 100  
## 15 Male Area.Income 0.00252 9.36e- 1 -0.0595 0.0645 100

library(corrplot)

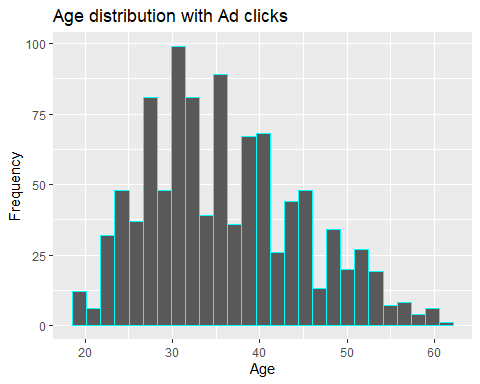
## corrplot 0.84 loaded

ads\_num <- Filter(is.numeric, ads)  
corrplot(cor(ads\_num))



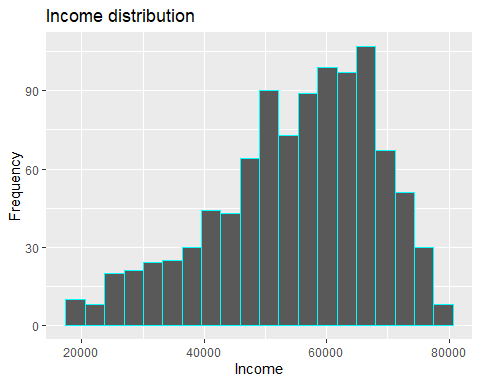
We plan on using the *Clicked.On.Ad* feature to determine fill colors for these graphs, but that won’t work if they stay as they’re currently set (integer data type).

library(ggplot2)  
  
ggplot(data = ads, aes(x = Age, fill = Clicked.on.Ad))+  
 geom\_histogram(bins = 27, color = 'cyan') +   
 labs(title = 'Age distribution with Ad clicks', x = 'Age', y = 'Frequency', fill = 'Clicked.on.Ad') +  
 scale\_color\_brewer(palette = 'Set2')

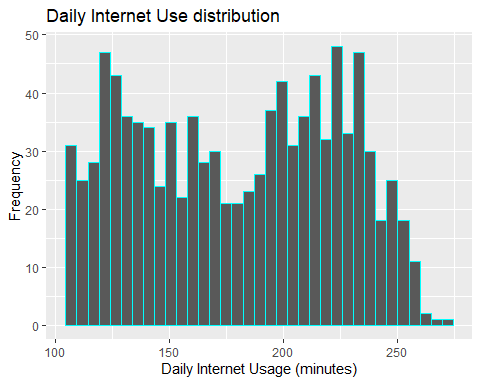


**Income and Click on Ad distribution**

ggplot(data = ads, aes(x = Area.Income, fill = Clicked.on.Ad))+  
 geom\_histogram(bins = 20, color = 'cyan') +   
 labs(title = 'Income distribution', x = 'Income', y = 'Frequency', fill = 'Clicked.on.Ad') +  
 scale\_color\_brewer(palette = 'Set1')

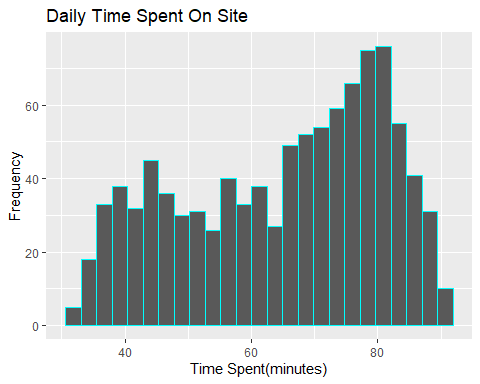
 **Daily Internet Use and the clicked on ad relationship**

ggplot(data = ads, aes(x = Daily.Internet.Usage, fill = Clicked.on.Ad))+  
 geom\_histogram(bins = 35, color = 'cyan') +   
 labs(title = 'Daily Internet Use distribution', x = 'Daily Internet Usage (minutes)', y = 'Frequency', fill = 'Clicked.on.Ad') +  
 scale\_color\_brewer(palette = 'Set1')



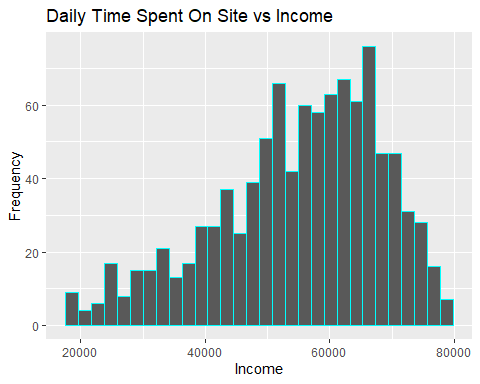
**Daily Time Spent on Site and the clicked on ad relationship**

ggplot(data = ads, aes(x = Daily.Time.Spent.on.Site, fill = Clicked.on.Ad))+  
 geom\_histogram(bins = 25, color = 'cyan') +   
 labs(title = 'Daily Time Spent On Site', x = 'Time Spent(minutes)', y = 'Frequency', fill = 'Clicked.on.Ad') +  
 scale\_color\_brewer(palette = 'Set1')



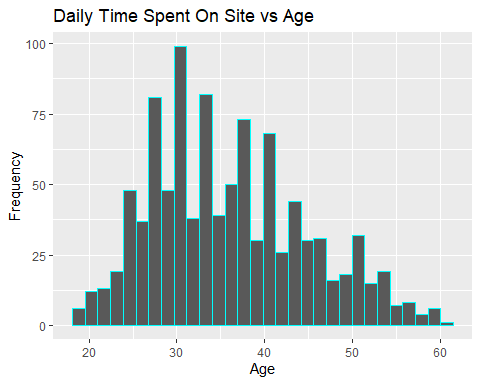
**Daily Time Spent on Site and the Income relationship**

ggplot(data = ads, aes(x =Area.Income , fill = Daily.Time.Spent.on.Site))+  
 geom\_histogram(bins = 30, color = 'cyan') +   
 labs(title = 'Daily Time Spent On Site vs Income', x = 'Income', y = 'Frequency', fill = 'Clicked.on.Ad') +  
 scale\_color\_brewer(palette = 'Set1')



**Daily Time Spent on Site and the Users’ Age relationship**

ggplot(data = ads, aes(x =Age , fill = Daily.Time.Spent.on.Site))+  
 geom\_histogram(bins = 30, color = 'cyan') +   
 labs(title = 'Daily Time Spent On Site vs Age', x = 'Age', y = 'Frequency', fill = 'Clicked.on.Ad') +  
 scale\_color\_brewer(palette = 'Set1')



**Daily Internet Usage per Country**

ads %>% group\_by(Country, Daily.Internet.Usage)%>% head(10)%>% arrange(desc(Daily.Internet.Usage))

## # A tibble: 10 x 10  
## # Groups: Country, Daily.Internet.Usage [10]  
## Daily.Time.Spen~ Age Area.Income Daily.Internet.~ Ad.Topic.Line City Male  
## <dbl> <int> <dbl> <dbl> <chr> <chr> <int>  
## 1 69.0 35 61834. 256. Cloned 5thge~ Wrig~ 0  
## 2 74.2 29 54806. 246. Triple-buffe~ West~ 1  
## 3 69.5 26 59786. 236. Organic bott~ Davi~ 0  
## 4 60.0 23 59762. 227. Sharable cli~ Jami~ 1  
## 5 68.4 35 73890. 226. Robust logis~ Sout~ 0  
## 6 74.5 30 68862 222. Configurable~ West~ 1  
## 7 88.9 33 53853. 208. Enhanced ded~ Bran~ 0  
## 8 80.2 31 68442. 194. Monitored na~ West~ 1  
## 9 69.9 20 55642. 184. Mandatory ho~ Rami~ 1  
## 10 66 48 24593. 132. Reactive loc~ Port~ 1  
## # ... with 3 more variables: Country <chr>, Timestamp <dttm>,  
## # Clicked.on.Ad <int>

**Daily Time Spent on Site per Country**

ads %>% group\_by(Country, Daily.Time.Spent.on.Site)%>% head(10)%>% arrange(desc(Daily.Time.Spent.on.Site))

## # A tibble: 10 x 10  
## # Groups: Country, Daily.Time.Spent.on.Site [10]  
## Daily.Time.Spen~ Age Area.Income Daily.Internet.~ Ad.Topic.Line City Male  
## <dbl> <int> <dbl> <dbl> <chr> <chr> <int>  
## 1 88.9 33 53853. 208. Enhanced ded~ Bran~ 0  
## 2 80.2 31 68442. 194. Monitored na~ West~ 1  
## 3 74.5 30 68862 222. Configurable~ West~ 1  
## 4 74.2 29 54806. 246. Triple-buffe~ West~ 1  
## 5 69.9 20 55642. 184. Mandatory ho~ Rami~ 1  
## 6 69.5 26 59786. 236. Organic bott~ Davi~ 0  
## 7 69.0 35 61834. 256. Cloned 5thge~ Wrig~ 0  
## 8 68.4 35 73890. 226. Robust logis~ Sout~ 0  
## 9 66 48 24593. 132. Reactive loc~ Port~ 1  
## 10 60.0 23 59762. 227. Sharable cli~ Jami~ 1  
## # ... with 3 more variables: Country <chr>, Timestamp <dttm>,  
## # Clicked.on.Ad <int>

People aged 33 seem to spend the most time on site therefore using the most internet.

#Conclusion 33year olds spend the most time on the site, generally people in their late 20’s till 33years old. People with higher income spend more time on the sight compared to the less income eraners

#Recommendation age had the highest value of a positive Pearson correlation. It is recommended that the client prioritizes people aged between 29-33 and bear this is mind as they strategise on marketing their products.

This project has been a success.