R - EDA

Lydia Gathoni

29 February 2020

## Research 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.

In order to work on the above problem, you need to do the following:

Define the question, the metric for success, the context, experimental design taken and the appropriateness of the available data to answer the given question

Find and deal with outliers, anomalies, and missing data within the dataset.

Perform univariate and bivariate analysis recording your observations.

Implement the solution by performing the respective analysis i.e. factor analysis, principal component analysis, and discriminant analysis.

Challenge your solution by providing insights on how you can make improvements.

adv <- read.csv("advertising.csv")  
head(adv)

## 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

### Explore the dataset

Summary, information, dimension

# summary of the dataset  
summary(adv)

## 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   
## Adaptive 24hour Graphic Interface : 1 Lisamouth : 3   
## Adaptive asynchronous attitude : 1 Williamsport : 3   
## Adaptive context-sensitive application : 1 Benjaminchester: 2   
## Adaptive contextually-based methodology: 1 East John : 2   
## Adaptive demand-driven knowledgebase : 1 East Timothy : 2   
## Adaptive uniform capability : 1 Johnstad : 2   
## (Other) :994 (Other) :986   
## Male Country Timestamp Clicked.on.Ad  
## Min. :0.000 Czech Republic: 9 2016-01-01 02:52:10: 1 Min. :0.0   
## 1st Qu.:0.000 France : 9 2016-01-01 03:35:35: 1 1st Qu.:0.0   
## Median :0.000 Afghanistan : 8 2016-01-01 05:31:22: 1 Median :0.5   
## Mean :0.481 Australia : 8 2016-01-01 08:27:06: 1 Mean :0.5   
## 3rd Qu.:1.000 Cyprus : 8 2016-01-01 15:14:24: 1 3rd Qu.:1.0   
## Max. :1.000 Greece : 8 2016-01-01 20:17:49: 1 Max. :1.0   
## (Other) :950 (Other) :994

# information about the dataset  
str(adv)

## '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 ...

# dimension of the dataset  
dim(adv)

## [1] 1000 10

The dataset has 1000 rows and 10 columns

### Data Cleaning

#### Missing Values

# Total missing values in each column   
# by using the function colSums()  
  
colSums(is.na(adv))

## 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 has no missing values

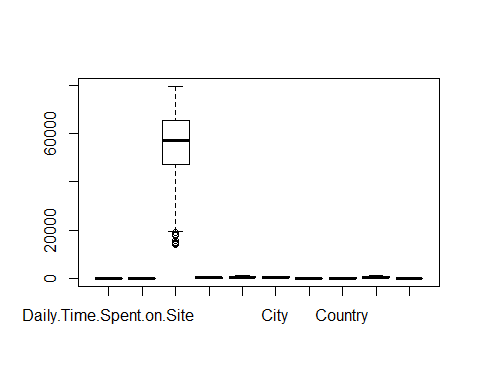
#### Duplicated rows

# duplicated rows in the dataset df   
# and assign to a variable duplicated\_rows  
  
duplicated\_rows <- adv[duplicated(adv),]  
  
# Lets print out the variable duplicated\_rows and see these duplicated rows   
  
duplicated\_rows

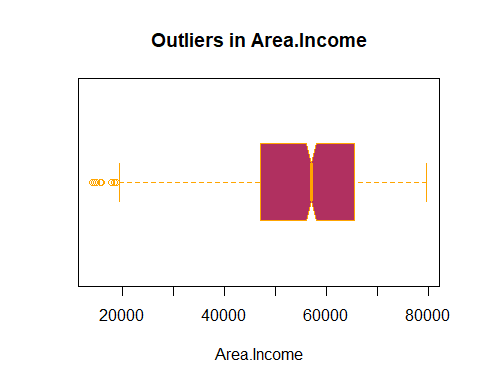
## [1] Daily.Time.Spent.on.Site Age Area.Income   
## [4] Daily.Internet.Usage Ad.Topic.Line City   
## [7] Male Country Timestamp   
## [10] Clicked.on.Ad   
## <0 rows> (or 0-length row.names)

We have no duplicated rows in the dataset

# plot a boxplot to help us visualise any existing outliers   
  
boxplot(adv  
 )

 We identify outliers in the Area.Income column We can narrow down to list out the outliers in that column

boxplot(adv$Area.Income,  
main = "Outliers in Area.Income",  
xlab = "Area.Income",  
col = "maroon",  
border = "orange",  
horizontal = TRUE,  
notch = TRUE  
)

 We identify outliers in the column

### Exploratory Data Analysis

#### Univariate Analysis

table(adv$Age)

##   
## 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44   
## 6 6 6 13 19 21 27 37 33 48 48 39 60 38 43 39 39 50 36 37 30 36 32 26 23 21   
## 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61   
## 30 18 13 16 18 20 12 15 10 9 7 2 6 4 2 4 1

Most of the people in the dataset are 31 years with 60 people

#### Mean daily time spent on site

# mean of the daily time spent on site  
mean(adv$Daily.Time.Spent.on.Site)

## [1] 65.0002

65 minutes is the average time spent on site

#### Mean daily internet usage

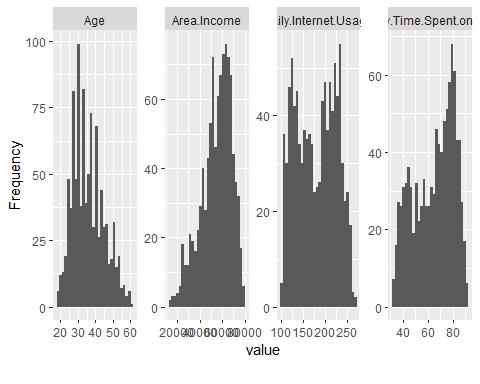
mean(adv$Daily.Internet.Usage)

## [1] 180.0001

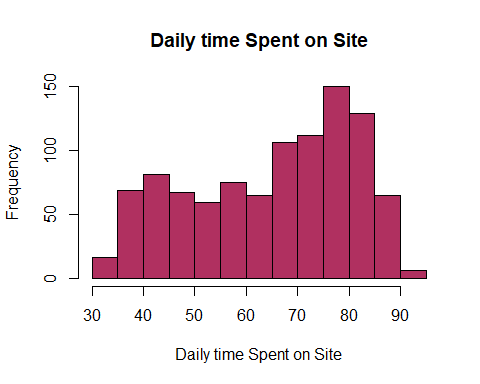
180 minutes is the average daily internet usage

##### Histograms

# Histograms  
library(DataExplorer)  
plot\_histogram(adv)

 ##### Daily Time spent on site

# Daily time spent on site distribution  
#  
x = hist(adv$Daily.Time.Spent.on.Site,  
 main = "Daily time Spent on Site",  
 xlab = "Daily time Spent on Site",  
 col = "maroon"  
)

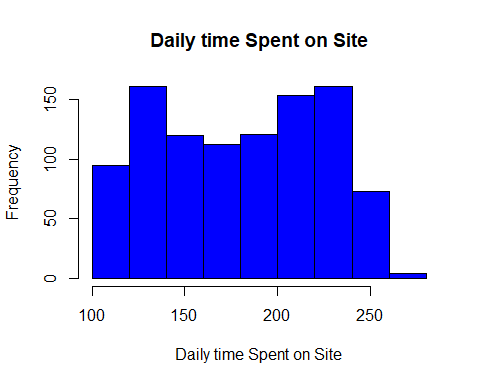
 The data is skewed to the right more people spend more time on the site

summary(adv$Daily.Time.Spent.on.Site)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 32.60 51.36 68.22 65.00 78.55 91.43

##### Daily Internet Usage

# Daily Internet Usage  
#  
y = hist(adv$Daily.Internet.Usage,  
 main = "Daily time Spent on Site",  
 xlab = "Daily time Spent on Site",  
 col = "blue"  
)

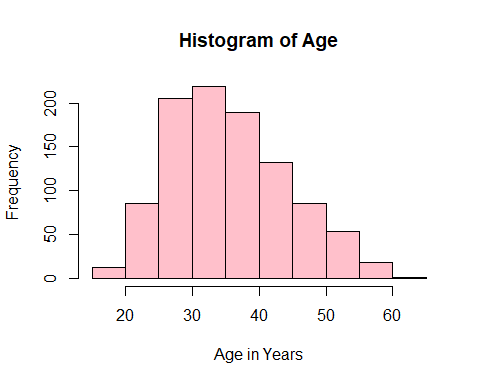


summary(adv$Daily.Internet.Usage)

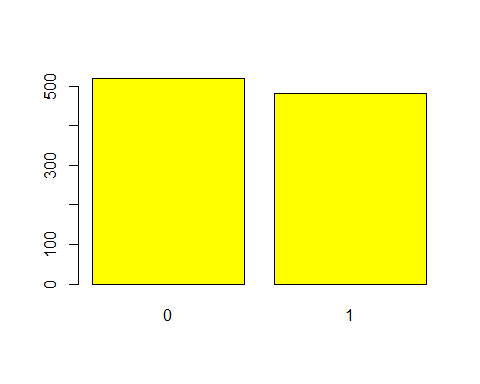
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 104.8 138.8 183.1 180.0 218.8 270.0

The average daily internet usage is 180 minutes

hist(adv$Age,  
 main = "Histogram of Age",  
 xlab = "Age in Years",  
 col = "pink")

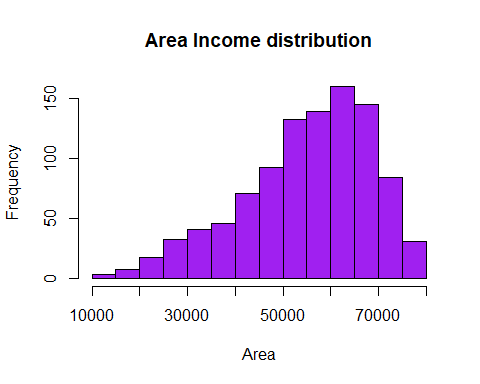
 Most people are between age 30 and 35 with the least being above 60

# fetch the frequency of gender from the dataset  
gender <- adv$Male  
gender\_freq <- table(gender)  
barplot(gender\_freq,  
 col = "yellow")

 The females represented by 0 are more than the males

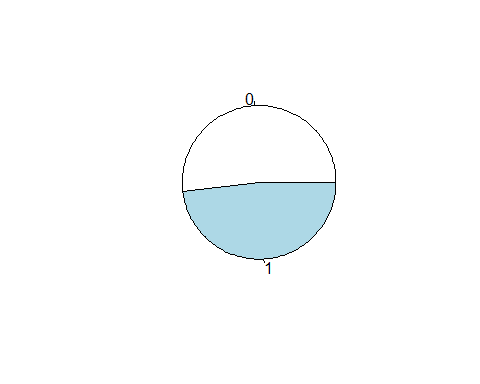
##### Distribution of Area income

z = hist(adv$Area.Income,  
 main = "Area Income distribution",  
 xlab = "Area",  
 col = "purple"  
)

 Area income is skewed to the right

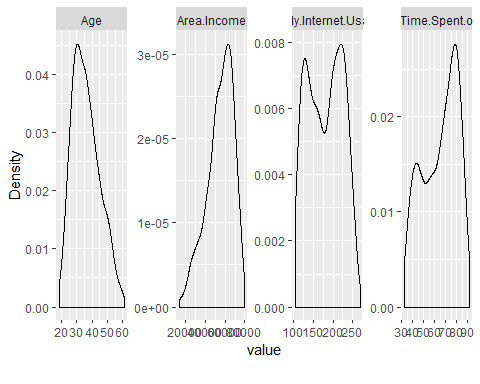
##### Pie Chart

library(DataExplorer)  
pie(table(adv$Male))

 0 value which represents females is more than 1 which represents males

##### Density plot

library(DataExplorer)  
plot\_density(adv)



#### Bivariate Analysis

# checking for covariance  
# covariance is a number that reflects the degree to which two variable vary together  
  
timespent <- adv$Daily.Time.Spent.on.Site  
internetusage<- adv$Daily.Internet.Usage  
  
# Using the cov() function to determine the covariance  
  
cov(timespent, internetusage)

## [1] 360.9919

A high covariance basically indicates there is a strong relationship between the variables We have a covariance of 360 which means this two are positively highly related

Correlation

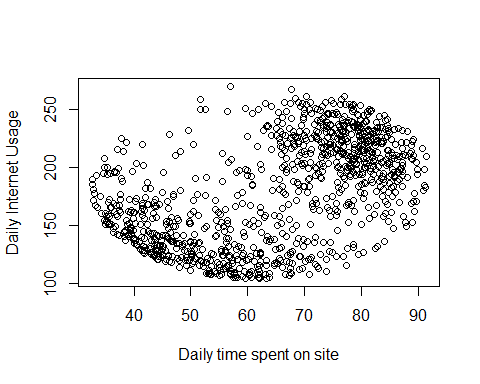
# checking for correlation  
# correlation is a normalized measurement of how the two are linearly related  
  
timespent <- adv$Daily.Time.Spent.on.Site  
internetusage<- adv$Daily.Internet.Usage  
  
# Using the cor() function to determine the covariance  
  
cor(timespent, internetusage)

## [1] 0.5186585

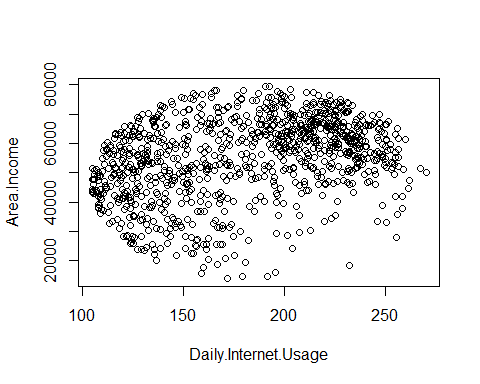
Shows a relation between the two

##### Scatterplot

# Scatterplot  
  
timespent <- adv$Daily.Time.Spent.on.Site  
internetusage<- adv$Daily.Internet.Usage  
  
plot(timespent, internetusage, xlab="Daily time spent on site", ylab="Daily Internet Usage")

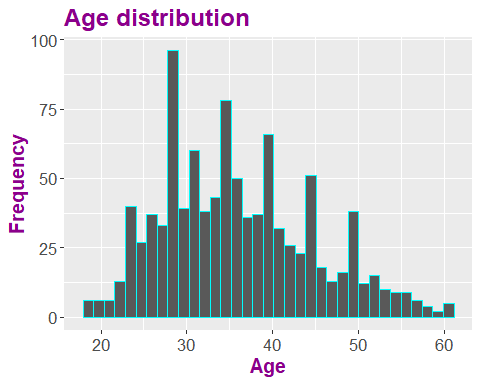


plot(Area.Income ~ Daily.Internet.Usage, data = adv)



##### Pairplot

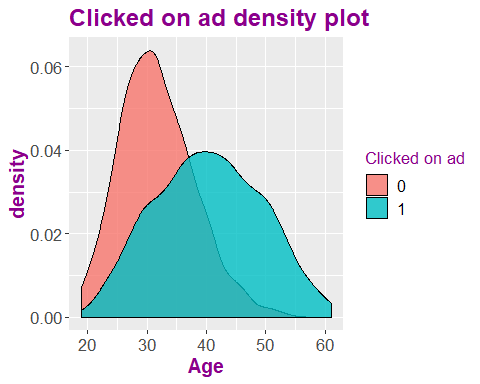
# Plotting a a pair of histograms  
library(ggplot2)  
options(repr.plot.width = 13, repr.plot.height = 7)  
ggplot(data = adv, aes(x = Age, fill = Clicked.on.Ad))+  
 geom\_histogram(bins = 35, color = 'cyan') +  
 labs(title = 'Age distribution', x = 'Age', y = 'Frequency', fill = 'Clicked on ad') +  
 scale\_color\_brewer(palette = 'Set1') +  
 theme(plot.title = element\_text(size = 18, face = 'bold', color = 'darkmagenta'),  
 axis.title.x = element\_text(size = 15, face = 'bold', color = 'darkmagenta'),  
 axis.title.y = element\_text(size = 15, face = 'bold', color = 'darkmagenta'),  
 axis.text.x = element\_text(size = 13, angle = 0),  
 axis.text.y = element\_text(size = 13),  
 legend.title = element\_text(size = 13, color = 'darkmagenta'),  
 legend.text = element\_text(size = 12))



# Plotting density plot  
library(ggplot2)  
options(repr.plot.width = 13, repr.plot.height = 7)  
p1 = ggplot(data = adv, aes(Age)) +  
 geom\_density(aes(fill=factor(Clicked.on.Ad)), alpha = 0.8) +  
 labs(title = 'Clicked on ad density plot', x = 'Age', fill = 'Clicked on ad') +  
 scale\_color\_brewer(palette = 'cool') +  
 theme(plot.title = element\_text(size = 18, face = 'bold', color = 'darkmagenta'),  
 axis.title.x = element\_text(size = 15, face = 'bold', color = 'darkmagenta'),  
 axis.title.y = element\_text(size = 15, face = 'bold', color = 'darkmagenta'),  
 axis.text.x = element\_text(size = 13, angle = 0),  
 axis.text.y = element\_text(size = 13),  
 legend.title = element\_text(size = 13, color = 'darkmagenta'),  
 legend.text = element\_text(size = 12))

## Warning in pal\_name(palette, type): Unknown palette cool

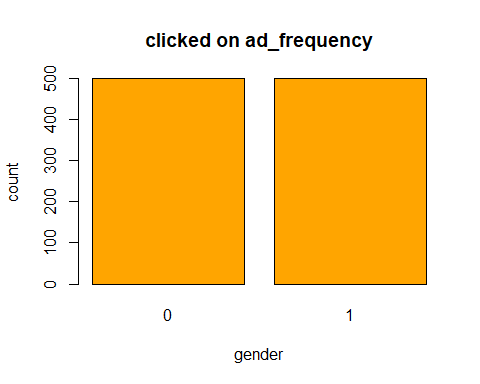
plot(p1)



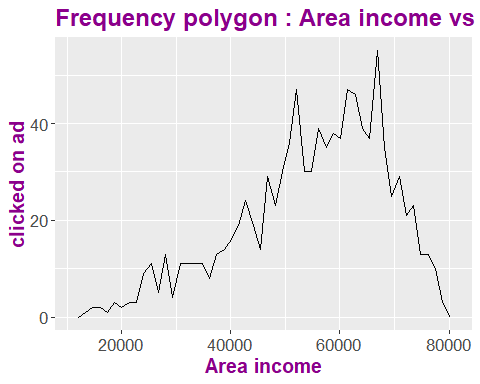
coad\_frequency = table(adv$Clicked.on.Ad)  
coad\_frequency

##   
## 0 1   
## 500 500

# Barplot for clicked on ad variable.  
#  
bar\_coad = barplot(coad\_frequency,  
 main = 'clicked on ad\_frequency',  
 xlab = 'gender',  
 ylab = 'count',  
 col = 'orange')

 The values 0 and 1 in the variable are even. This is a perrfectly balanced dataset.

# Frequency polygon  
library(ggplot2)  
options(repr.plot.width = 13, repr.plot.height = 7)  
ggplot(data = adv, aes(x = Area.Income, col = Clicked.on.Ad))+  
 geom\_freqpoly(bins = 50)+  
 labs(title = 'Frequency polygon : Area income vs clicked on ad', x = 'Area income', y = 'clicked on ad', fill = 'Clicked on ad') +  
 scale\_color\_brewer(palette = 'Set1') +  
 theme(plot.title = element\_text(size = 18, face = 'bold', color = 'darkmagenta'),  
 axis.title.x = element\_text(size = 15, face = 'bold', color = 'darkmagenta'),  
 axis.title.y = element\_text(size = 15, face = 'bold', color = 'darkmagenta'),  
 axis.text.x = element\_text(size = 13),  
 axis.text.y = element\_text(size = 13),  
 legend.title = element\_text(size = 13, color = 'darkmagenta'),  
 legend.text = element\_text(size = 12))



#### Multivariate Analysis

##### Correlation Plot

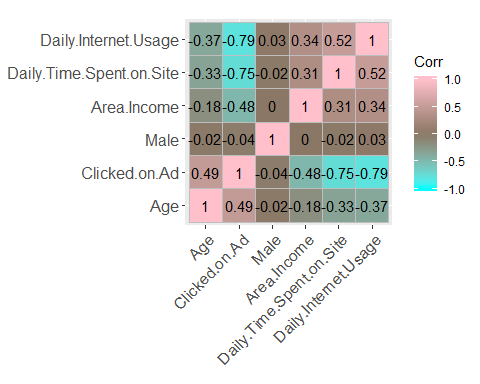
library(ggcorrplot)  
library(dplyr)

##   
## Attaching package: 'dplyr'

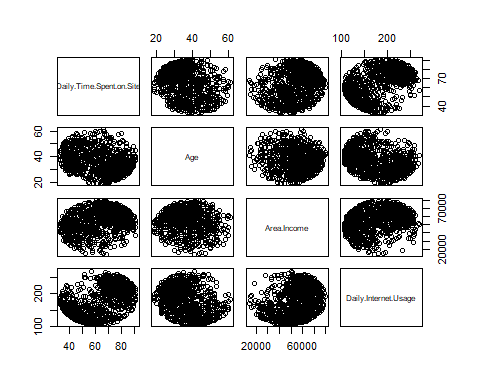
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

corr = round(cor(select\_if(adv, is.numeric)), 4)  
ggcorrplot(corr, hc.order = T, ggtheme = ggplot2::theme\_grey,  
 colors = c("cyan", "peachpuff4", "pink"), lab = T)

 We observe that daily time spent on site and daily internet usage are highly related

# Pairplot  
pairs(adv[,c(1,2,3,4)])



Most variables are related positively