Advertisement Analysis — Week 12 IP

1. Define the question

1.1 Main question —which individuals are more likely to click on cryptpgraphy course advertisement?

1.2 Metric for Success

1.3 Understanding the context

A Kenyan entrepreneur who has previously ran a course ad on her blog would like to know, which kinds of individuals are more likely to click on a new ad about Cryptography course. From the initial advertisement, the entrepreneur was able to collect some data. As Data Science Consultant, you have been tasked to analyze the given data and report recommendations to the entrepreneur. ### 1.4 Experimental Design The approach for this project will include: 1. Importing libraries 2. Reading the data 3. Checking the data 4. Tidying up the data 5. Implementing the solution using Univariate and Bivariate 6. Conclusion and Recommendations ### 1.5 Data Relevance

1. Importing libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts -----------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

2. Reading the data

```
# Imported data from the directory where it was saved
advertising <- read.csv("~/Moringa School/R Programming/R datasets/advertising.csv")
view(advertising)
```

3. Checking the data

```
# let's get glimpse of how the data looks like
glimpse(advertising)
```

```
## Rows: 1,000
## Columns: 10
## $ Daily.Time.Spent.on.Site <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, ...
## $ Age                      <int> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49...
## $ Area.Income              <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 73...
```

```
## $ Daily.Internet.Usage     <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 22...
## $ Ad.Topic.Line            <chr> "Cloned 5thgeneration orchestration", "Mon...
## $ City                     <chr> "Wrightburgh", "West Jodi", "Davidton", "W...
## $ Male                     <int> 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, ...
## $ Country                  <chr> "Tunisia", "Nauru", "San Marino", "Italy",...
## $ Timestamp                <chr> "2016-03-27 00:53:11", "2016-04-04 01:39:0...
## $ Clicked.on.Ad            <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, ...
```

We see that our data has 1,000 records and 10 columns. Some columns have integer, float and character data types. clicked.on.ad is our target variable.

```
#checking the top of the data
head(advertising)
```

```
##   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 bottom of the data
tail(advertising)
```

```
##      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
```

After looking at the head and tail of the data, there are a few assumptions on the data we need to make before we proceed. These assumptions include; 1. Daily time spend on site column is in minutes 2. Income is in United States Dollars($) 3. Daily internet usage units is in megabytes. Also assuming that the megabytes refers to data used on the blog site 4. For male column if 0 that means not male, but if 1 then yes it's a male 5. Clicked on ad column: if 1 that means yes somebody clicked on the ad, but if 0 then the add was not clicked

```
#checking for statistical summaries of numerical data
#first select numerical variables
num <- select(advertising, Daily.Time.Spent.on.Site, Age, Area.Income,Daily.Internet.Usage)
glimpse(num)
```

```
## Rows: 1,000
## Columns: 4
## $ Daily.Time.Spent.on.Site <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, ...
## $ Age                      <int> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49...
## $ Area.Income              <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 73...
## $ Daily.Internet.Usage     <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 22...
```

```
summary(num)
```

```
##  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
```

On average daily time spent on the site(her blog) is 65 minutes Average age that visit the site is about 36 years old Average income area is $55,000 The average daily internet usage on the site is 180 megabytes From the statistical summary, we don't see anything strange in the numerical columns

```
# checking for unique elements in columns if interest
# starting with our target variable
unique(advertising$Clicked.on.Ad)
```

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

There are two elements 0 for not clicked and 1 for clicked(based on the assumption we made above)

```
unique(advertising$Male)
```

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

There is two elements in the male column. 0 for not a male and 1 for male

```
#unique(advertising$Country)
```

In total there is a total of 237 countries. From the countries list, I see funny country names e.g Nauru, Wallis & Futuna, Holy see. Where did these countries come from? Note: country output deleted because the list is too long

4. Tidying the data

4.1 Missing values

```
# checking for missing values in each column
colSums(is.na(advertising))
```

```
## 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

4.2 Duplicates

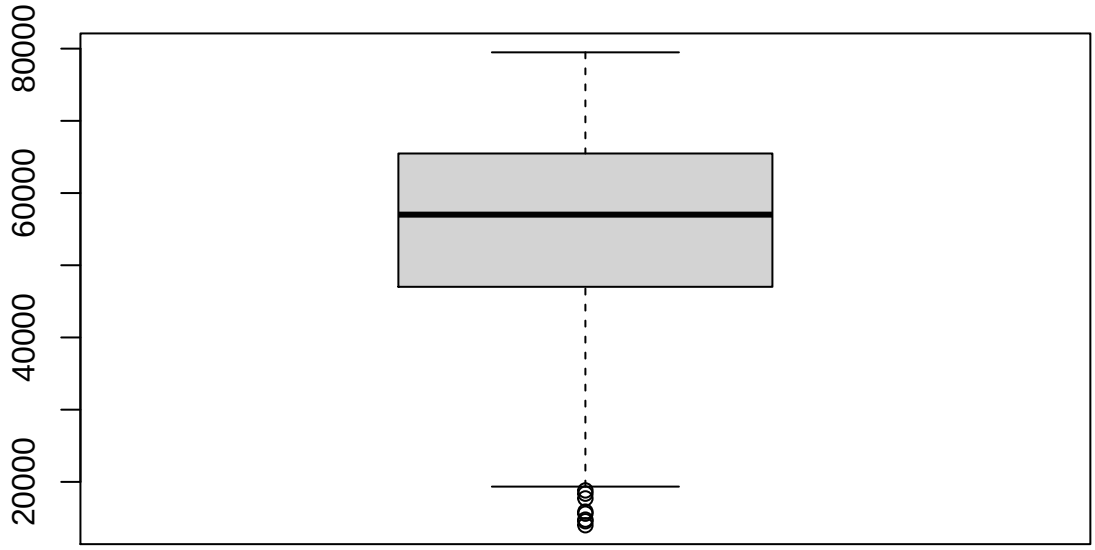
```
# checking for duplicates
advert_duplicates <- advertising[duplicated(advertising),]
advert_duplicates
```

```
##  [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)
```

There are no duplicates in our dataset

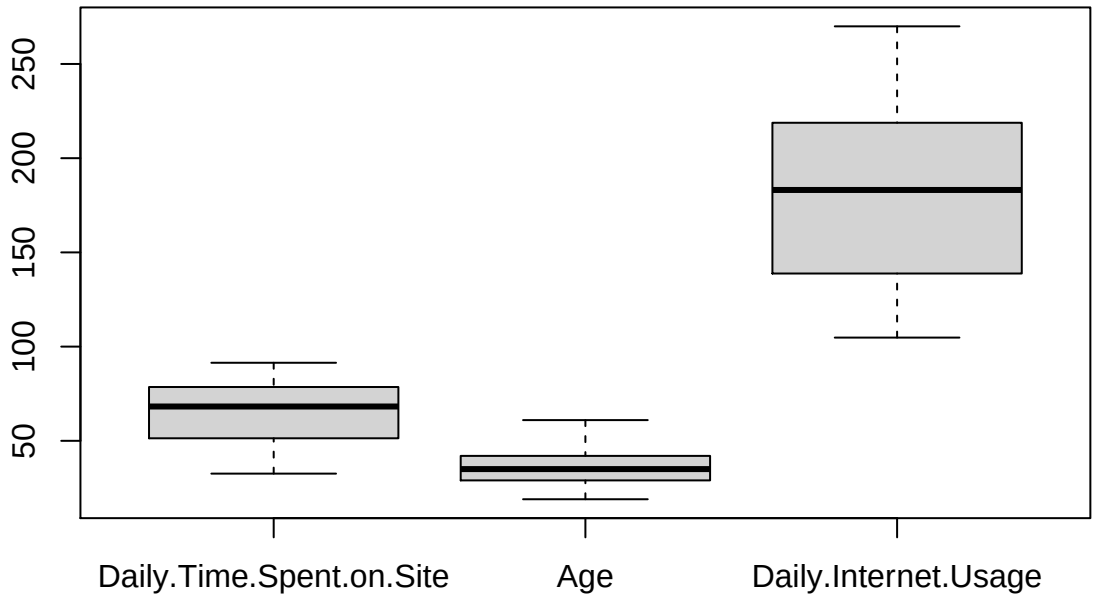
4.3 Outliers

```
# checking for outliers in numerical variables using boxplots
# we will separate area income from other numerical variables because of the difference in scale
boxplot(advertising$Area.Income)
```



We see some outliers from Income column on the lower side of the whisker. We won't be deleting these outliers because we establish that these are true observations. The low income could be as a result of many reasons. An explanation for these observations is that — in every country there are certain areas that have a lot of poor people, which leads to low area income.

```
# checking for outliers in the other numerical columns
num_cols <- select(advertising, Daily.Time.Spent.on.Site, Age, Daily.Internet.Usage)
boxplot(num_cols)
```



There are no observed outliers in these columns. Based on the analysis, so far, our data is pretty clean and we are ready to begin solution implementation using Univariate and Bivariate analysis.

5. Implementing the solution using Univariate and Bivariate

5.1 Univariate analysis

1. Measures of central tendency and measure of dispersion for numerical variables

```
# finding the mean
colMeans(num)
```

```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##                  65.0002                  36.0090              55000.0001
##     Daily.Internet.Usage
##                180.0001
```

```
# finding the median, min, max, range, quantiles using statistical summary code
summary(num)
```

```
##  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
```

```
# Variance
# Daily.Time.Spent.on.Site column
sapply(num, var)
```

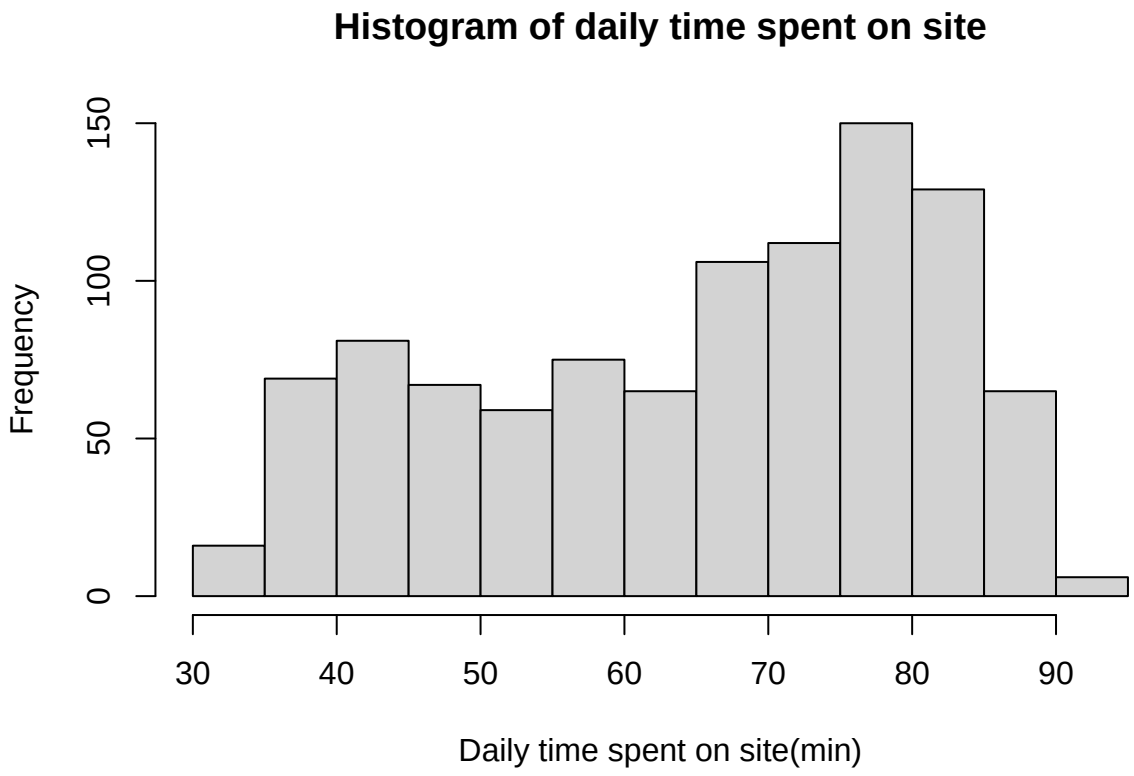
```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##             2.513371e+02             7.718611e+01             1.799524e+08
##     Daily.Internet.Usage
##             1.927415e+03
```

```
# standard deviation
sapply(num, sd)
```

```
## Daily.Time.Spent.on.Site                      Age             Area.Income
##                15.853615                 8.785562             13414.634022
##     Daily.Internet.Usage
##                43.902339
```

2. Histograms

```
hist(num$Daily.Time.Spent.on.Site, xlab= "Daily time spent on site(min)", ylab="Frequency",
     main="Histogram of daily time spent on site")
```



From the graph, we see that the most popular times spent on the site is between 65-85 minutes.