#WEEK 12 R INDEPENDENT PROJECT #By: Brian Onchweri #DSC14

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

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

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```r
advertising <- fread('http://bit.ly/IPAdvertisingData')
```

```r
#Creating a DataFrame out of the CSV File we loaded

df <- data.frame((advertising))

#Previewing our new DataFrame
head(advertising,n=5)
```

```
##    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
##                             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
##              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
```

#Checking and Treating Duplicated Data

```
duplicated_rows <- df[duplicated(df),]


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 Duplicates as can be seen below
```

#Checking and Treating Missing Values

```
sum(is.na(df))
```

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

```
#Checking For missing values in our Dataset
colSums(is.na(df))
```

```
## 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
```

#Finding and Dealing with outliers

```
library(dplyr)
library(ggplot2)

my_custom_df<- as_tibble(df)
my_custom_df
```

```
## # A tibble: 1,000 x 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             80.2    31      68442.             194. Monitored na~ West~     1
##  3             69.5    26      59786.             236. Organic bott~ Davi~     0
##  4             74.2    29      54806.             246. Triple-buffe~ West~     1
##  5             68.4    35      73890.             226. Robust logis~ Sout~     0
##  6             60.0    23      59762.             227. Sharable cli~ Jami~     1
##  7             88.9    33      53853.             208. Enhanced ded~ Bran~     0
##  8             66      48      24593.             132. Reactive loc~ Port~     1
##  9             74.5    30      68862              222. Configurable~ West~     1
## 10             69.9    20      55642.             184. Mandatory ho~ Rami~     1
## # ... with 990 more rows, and 3 more variables: Country <chr>,
## #   Timestamp <dttm>, Clicked.on.Ad <int>
```

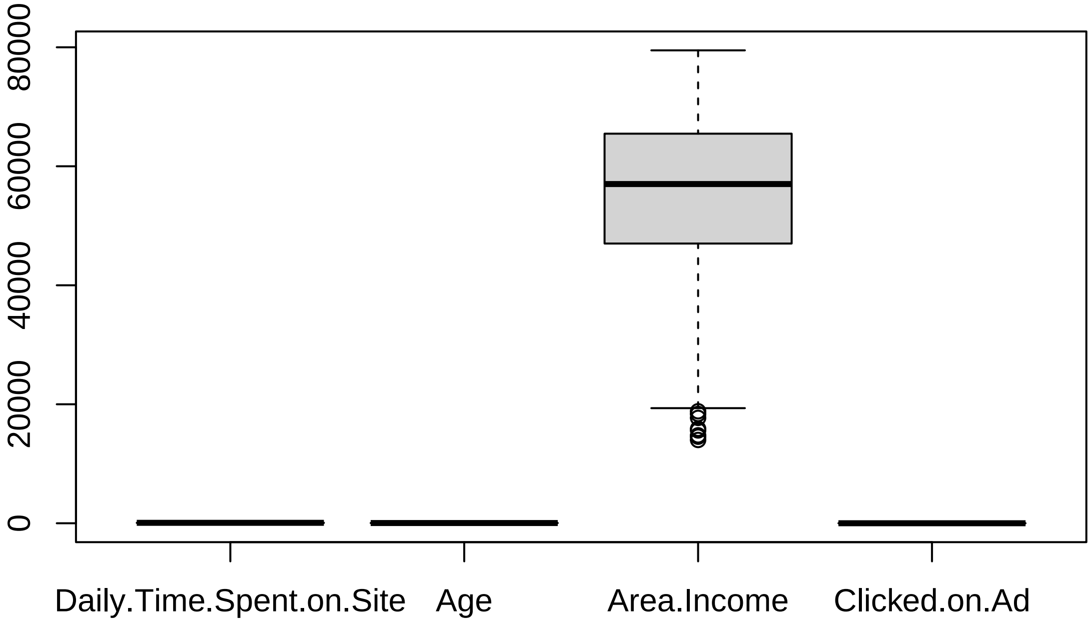
#Creating a custom Dataframe with only selected columns

```
my_custom_df<-my_custom_df%>% select(0,1,2,3,10)


my_custom_df
```

```
## # A tibble: 1,000 x 4
##    Daily.Time.Spent.on.Site   Age Area.Income Clicked.on.Ad
##                       <dbl> <int>       <dbl>         <int>
##  1                     69.0    35      61834.             0
##  2                     80.2    31      68442.             0
##  3                     69.5    26      59786.             0
##  4                     74.2    29      54806.             0
##  5                     68.4    35      73890.             0
##  6                     60.0    23      59762.             0
##  7                     88.9    33      53853.             0
##  8                     66      48      24593.             1
##  9                     74.5    30      68862              0
## 10                     69.9    20      55642.             0
## # ... with 990 more rows
```

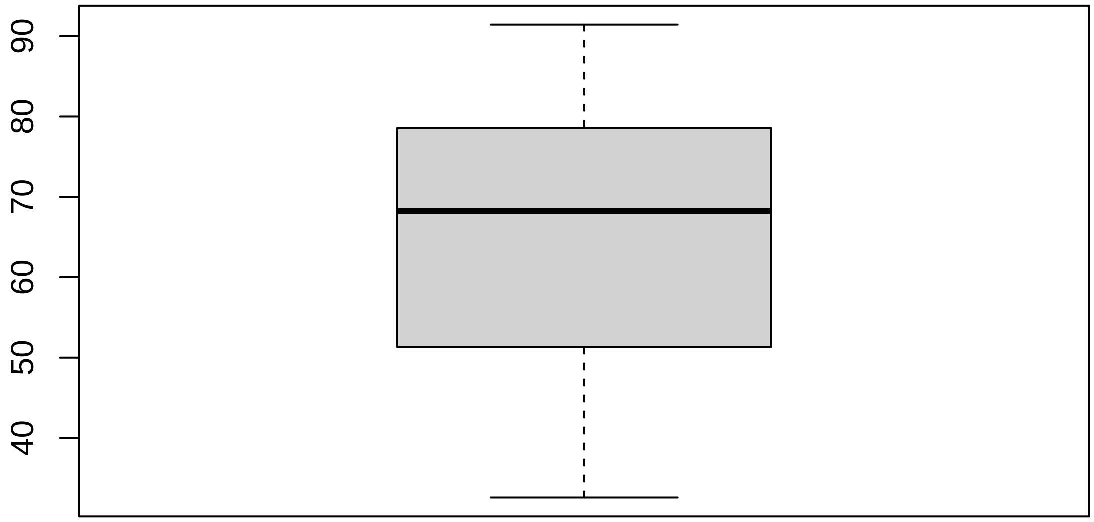
#UNIVARIATE AND MULTIVARIATE DATA ANALYSIS

```
#Converting dataframe items to numeric
my_custom_df<-lapply(my_custom_df,as.numeric)



boxplot(my_custom_df)
```



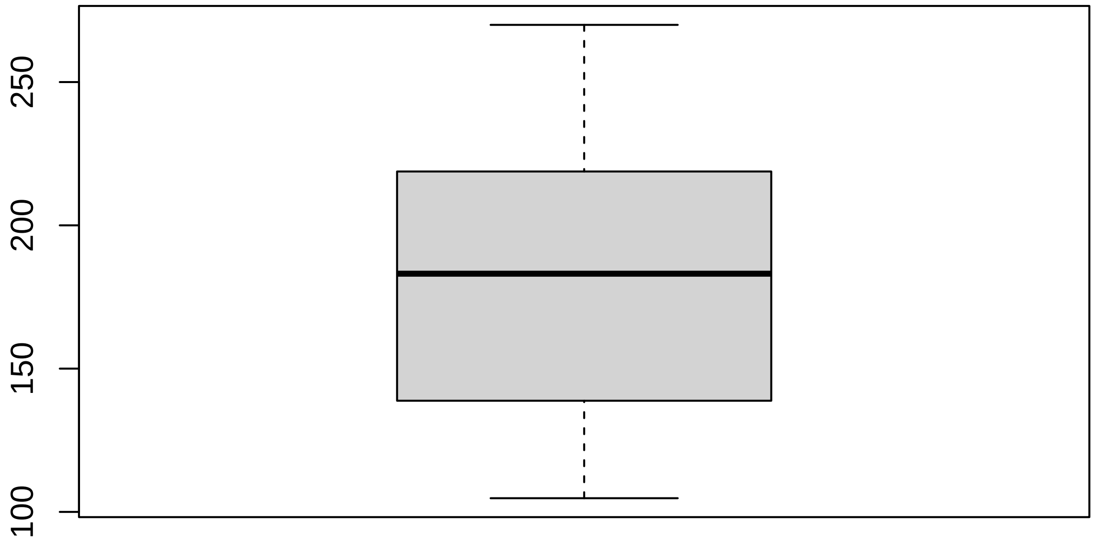
#Boxplot showing time spent on site

```
boxplot(df$Daily.Time.Spent.on.Site)
```



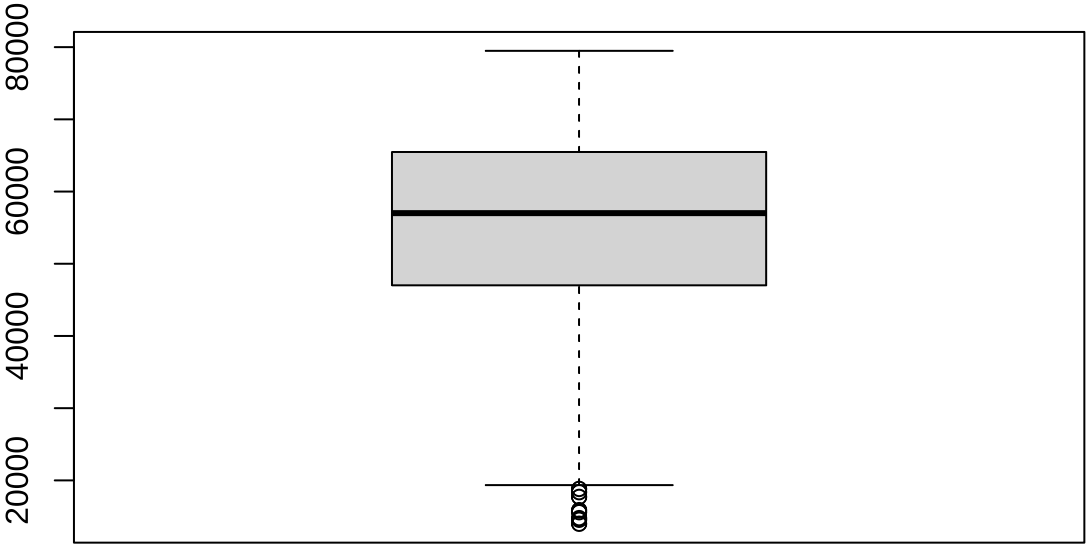
#Boxplot showing internet usage

```
boxplot(df$Daily.Internet.Usage)
```



#Boxplot showing Area Income

```
boxplot(df$Area.Income)
```



#We can Definitely see outliers in "Area.Income" column this warrants further investigation

```
cor(df$Area.Income, df$Clicked.on.Ad)
```

```
## [1] -0.4762546
```

#Correlation between time spent on site and AD CLICKED

```
cor(df$Daily.Time.Spent.on.Site,df$Clicked.on.Ad)
```

```
## [1] -0.7481166
```

#Correlation between Age and AD CLICKED

```
cor(df$Age,df$Clicked.on.Ad)
```

```
## [1] 0.4925313
```

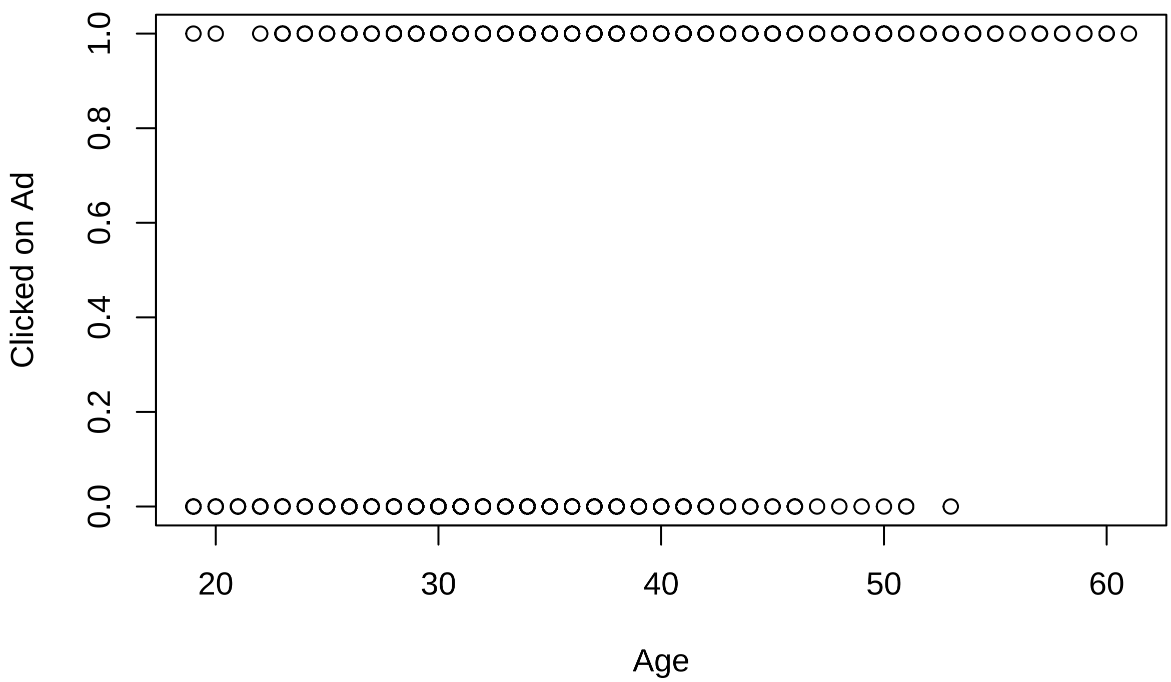
#Correlation between Internet usage and AD CLICKED

```
cor(df$Daily.Internet.Usage,df$Clicked.on.Ad)
```

```
## [1] -0.7865392
```

#Scatter plot showing Age and Ad clicked

```
plot(df$Age, df$Clicked.on.Ad, xlab="Age", ylab="Clicked on Ad")
```



#Creating a custom Function for Getting the 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)))]
                       }
```

```
# Calculating the mode using out getmode() function
# ---
#
Customer.clicks.mode <- getmode(df$Clicked.on.Ad)

Customer.clicks.mode
```

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

#We notice above that most people were not clicking the Ads. This could be useful information to our client but for now we will drop all the people who dint click on Ads and focus on the ones who did

```
maximum_clicks<-max(df$Clicked.on.Ad)
maximum_clicks
```

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

#To answer our research question we are going to subset our original Dataframe and remain with only the rows where Ads were clicked. Ideally this are our target customers we need to study since they are the ones who showed interest on the Ads, then we will go ahead and use our custom mode fxn to further subset the dataframe to narrow down on the customer we want to target our Ads at

```
customers_with_interest<-subset(df,Clicked.on.Ad!=0)
```

#Target Age for our Ads

```
getmode(customers_with_interest$Age)
```

```
## [1] 45
```

#Target Area Income for our Ads

```
getmode(customers_with_interest$Area.Income)
```

```
## [1] 24593.33
```

#Target Topic Line for our Ads

```
getmode(customers_with_interest$Ad.Topic.Line)
```

```
## [1] "Reactive local challenge"
```

#Target City

```
getmode(customers_with_interest$City)
```

```
## [1] "Robertfurt"
```

#Target Country

```
getmode(customers_with_interest$Country)
```

```
## [1] "Australia"
```

#Target timing for our Ads

```
getmode(customers_with_interest$Timestamp)
```

```
## [1] "2016-03-07 01:40:15 UTC"
```