ONLINE CRYPTOGRAPHY COURSE

VICTOR NYARIBO

5/19/2021

a) Data Analytic Question

The aim of this project is to identify individuals most likely to click on an online cryptography course advert.

b) Success Metrics

- Successful Loading the data.
- Successful Handling missing data.
- Successful Outliers detection.
- Successful Outlier Visualization.
- Successful Handling outliers.
- Successful Univariate analysis.
- Successful Bivariate analysis.

c) Context

Internet has become the most prominent and accessible way to spread the news about an event or to pitch, advertise and sell a product, globally. The success of any advertisement campaign lies in reaching the right class of target audience and eventually convert them as potential customers in the future. Businesses are predominantly charged based the number of clicks that they received for their advertisement while some websites also bill them with a fixed charge per billing cycle. This creates a necessity for the advertising firms to analyze and study these influential factors to achieve the maximum possible gain through the advertisements. Additionally, it is equally important for the businesses to customize these factors rightly to achieve the maximum clicks.

d) Data Understanding

Variables

- Daily Time Spent on a Site: Time spent by the user on a site in minutes.
- Age:Customer's age in terms of years.
- Area Income: Average income of geographical area of consumer.
- Daily Internet Usage: Average minutes in a day consumer is on the internet.
- Ad Topic Line: Headline of the advertisement.
- City: City of the consumer.
- Male: Whether or not a consumer was male.
- Country: Country of the consumer.
- Timestamp: Time at which user clicked on an Ad or the closed window.
- Clicked on Ad: 0 or 1 is indicated clicking on an Ad.

e) Experimental Design

- Formulation of the research question.
- Loading the data.

- Exploratory Data Analysis.
- Solution Implementation.
- Challenging the solution.
- Follow up .

Data Importation

```
advertising<-df <- read.csv("http://bit.ly/IPAdvertisingData",header =T)
```

converting data.frame data into data.table

```
advertising<-as.data.table(advertising)
class(advertising) #checking class
```

```
## [1] "data.table" "data.frame"
```

Data Columns

```
#advertising%>%head(2)
kable(colnames(advertising))
```

x
Daily.Time.Spent.on.Site
Age
Area.Income
Daily.Internet.Usage
Ad.Topic.Line
City
Male
Country
Timestamp
Clicked.on.Ad

Delete unneccessary columns

x
Daily.Time.Spent.on.Site
Age
Area.Income
Daily.Internet.Usage
City
Male
Country
Clicked.on.Ad

Check for missing values

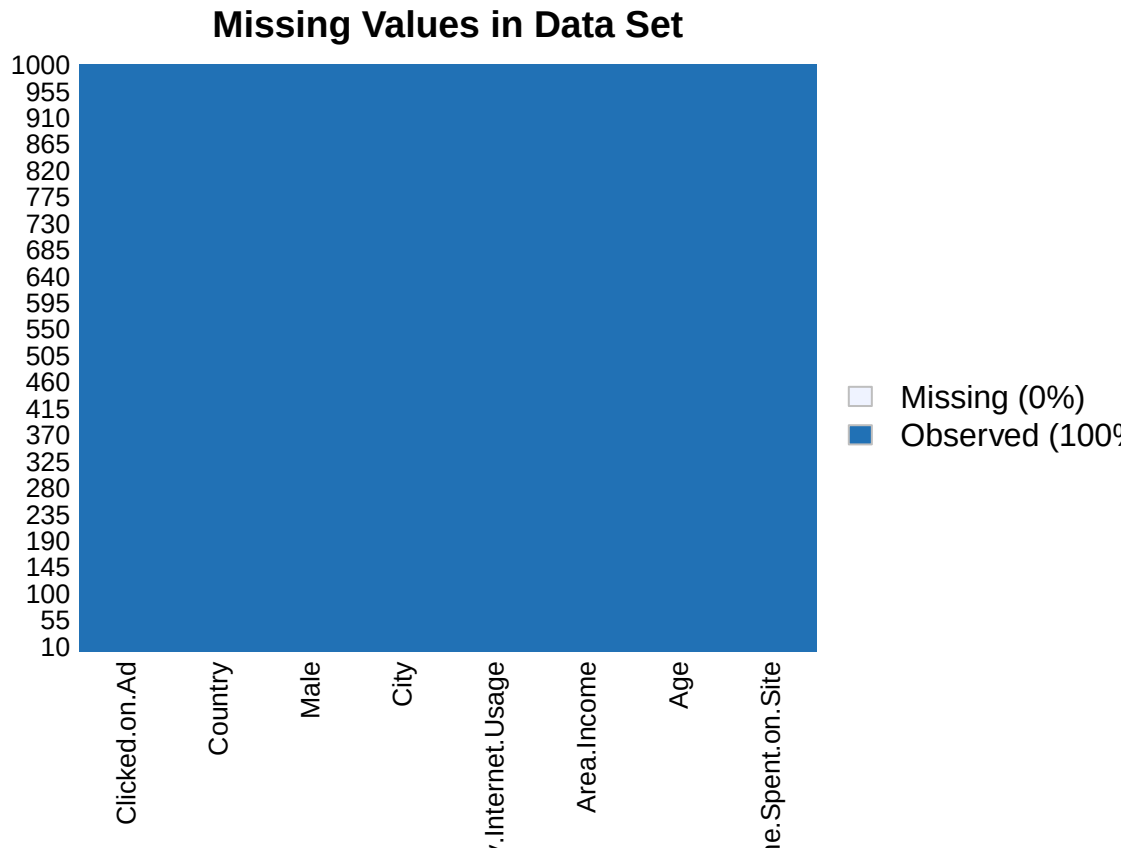
```
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 4.0.5
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.6, built: 2019-11-24)
## ## Copyright (C) 2005-2021 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
missmap(advertising,main="Missing Values in Data Set")
```



Tibbles

A tibble is a special kind of data.frame used by dplyr and other packages of the tidyverse. Tidyverse is a set of packages for data science that work in harmony because they share common data representations and API design. When a data.frame is turned into a tibble its class will change.

```
class(advertising)
```

```
## [1] "data.table" "data.frame"
```

```
advertising <- tbl_df(advertising)
```

```
## Warning: `tbl_df()` is deprecated as of dplyr 1.0.0.
## Please use `tibble::as_tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
class(advertising)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

Data Overview

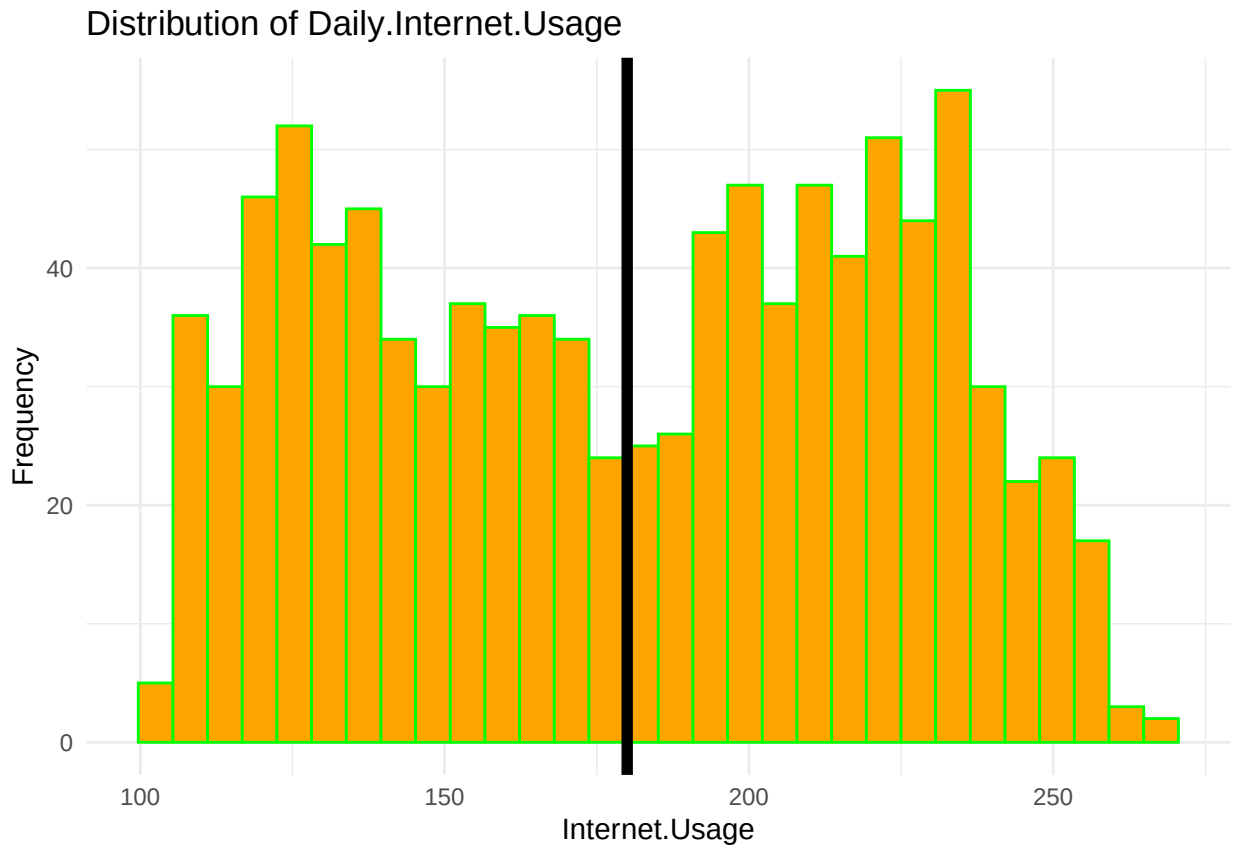
```
## Rows: 1,000
## Columns: 8
## $ Daily.Time.Spent.on.Site <dbl> 68.95, 80.23, 69.47, 74.15, 68.37, 59.99, ...
## $ Age                      <int> 35, 31, 26, 29, 35, 23, 33, 48, 30, 20, 49...
## $ Area.Income              <dbl> 61833.90, 68441.85, 59785.94, 54806.18, 73...
## $ Daily.Internet.Usage     <dbl> 256.09, 193.77, 236.50, 245.89, 225.58, 22...
## $ 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",...
## $ Clicked.on.Ad            <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, ...
```

Data preview

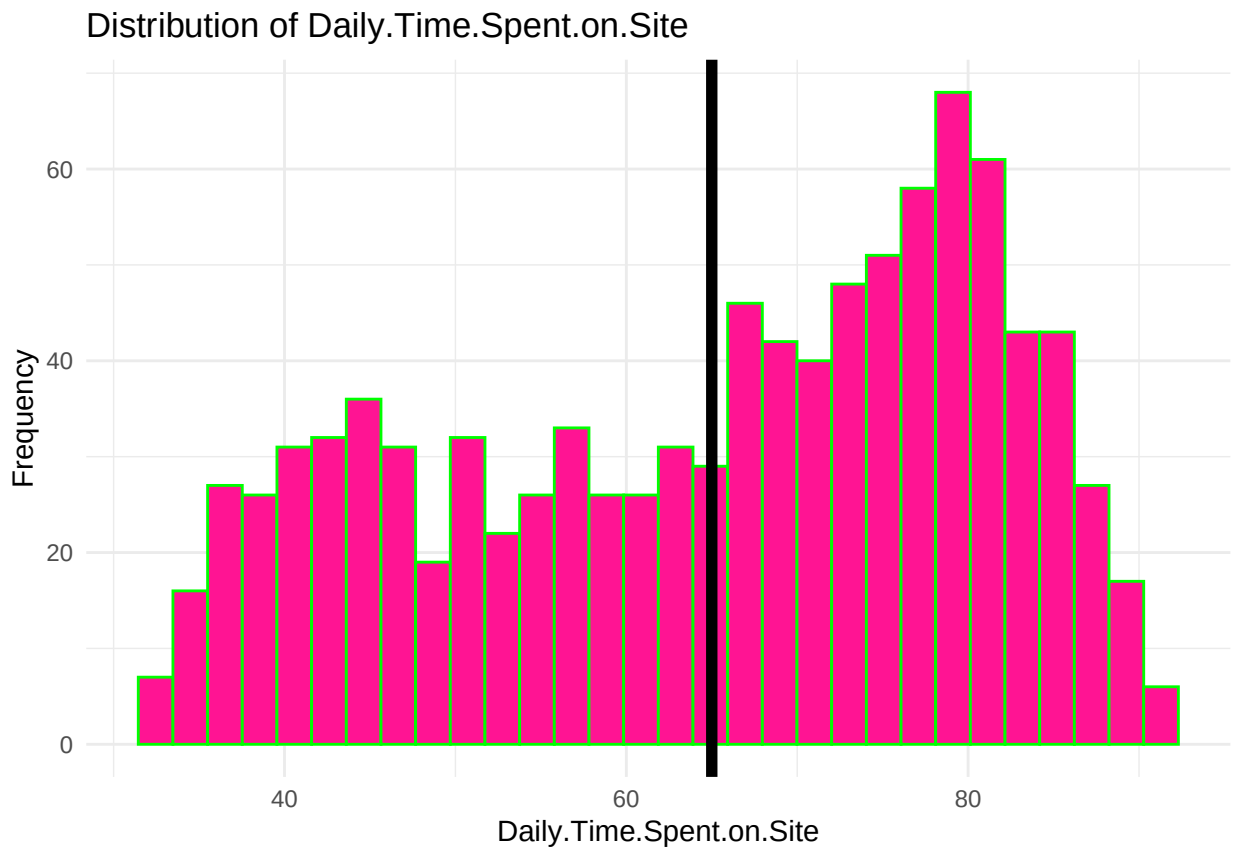
```
## # A tibble: 6 x 8
##   Daily.Time.Spen~   Age Area.Income Daily.Internet.~ City   Male Country
##              <dbl> <int>       <dbl>            <dbl> <chr> <int> <chr>
## 1             69.0    35      61834.             256. Wrig~     0 Tunisia
## 2             80.2    31      68442.             194. West~     1 Nauru
## 3             69.5    26      59786.             236. Davi~     0 San Ma~
## 4             74.2    29      54806.             246. West~     1 Italy
## 5             68.4    35      73890.             226. Sout~     0 Iceland
## 6             60.0    23      59762.             227. Jami~     1 Norway
## # ... with 1 more variable: Clicked.on.Ad <int>
```

Univariate analysis of a continuous variables

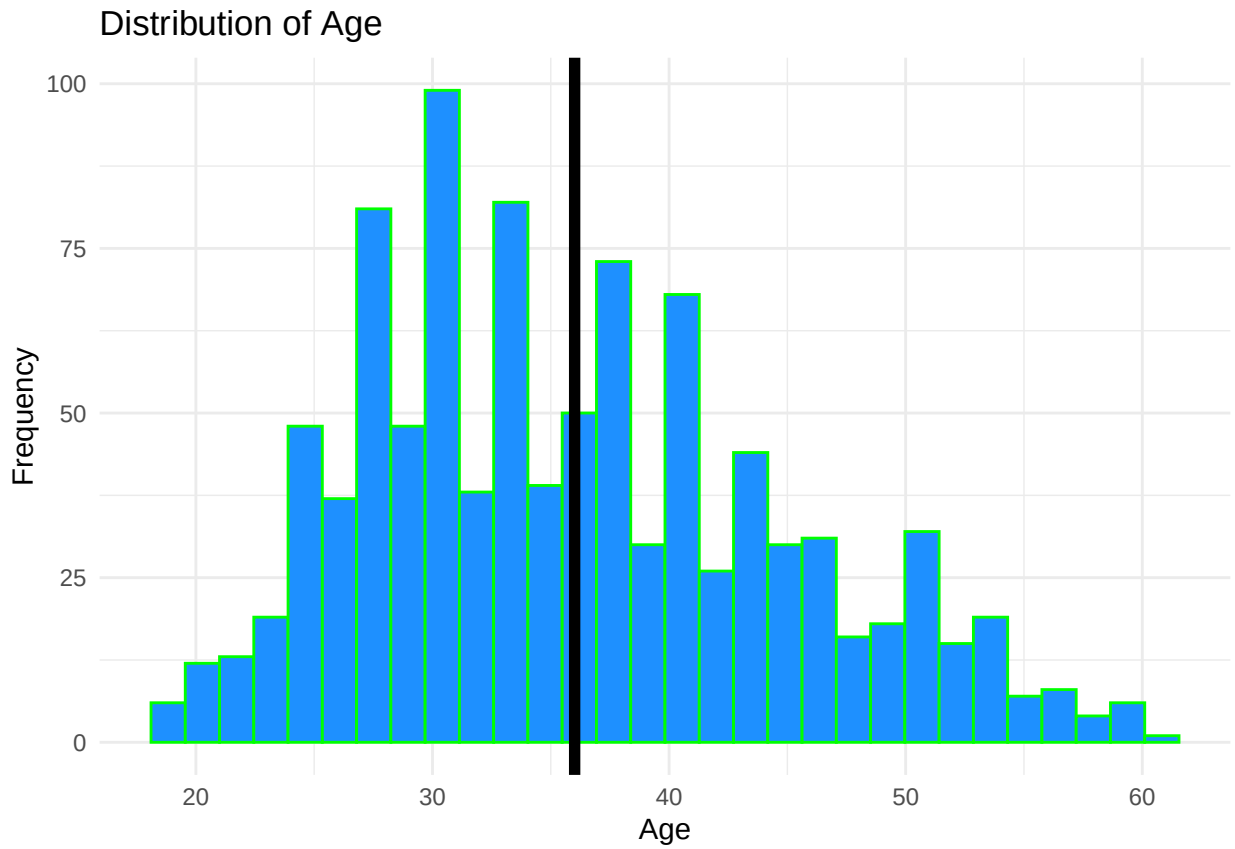
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



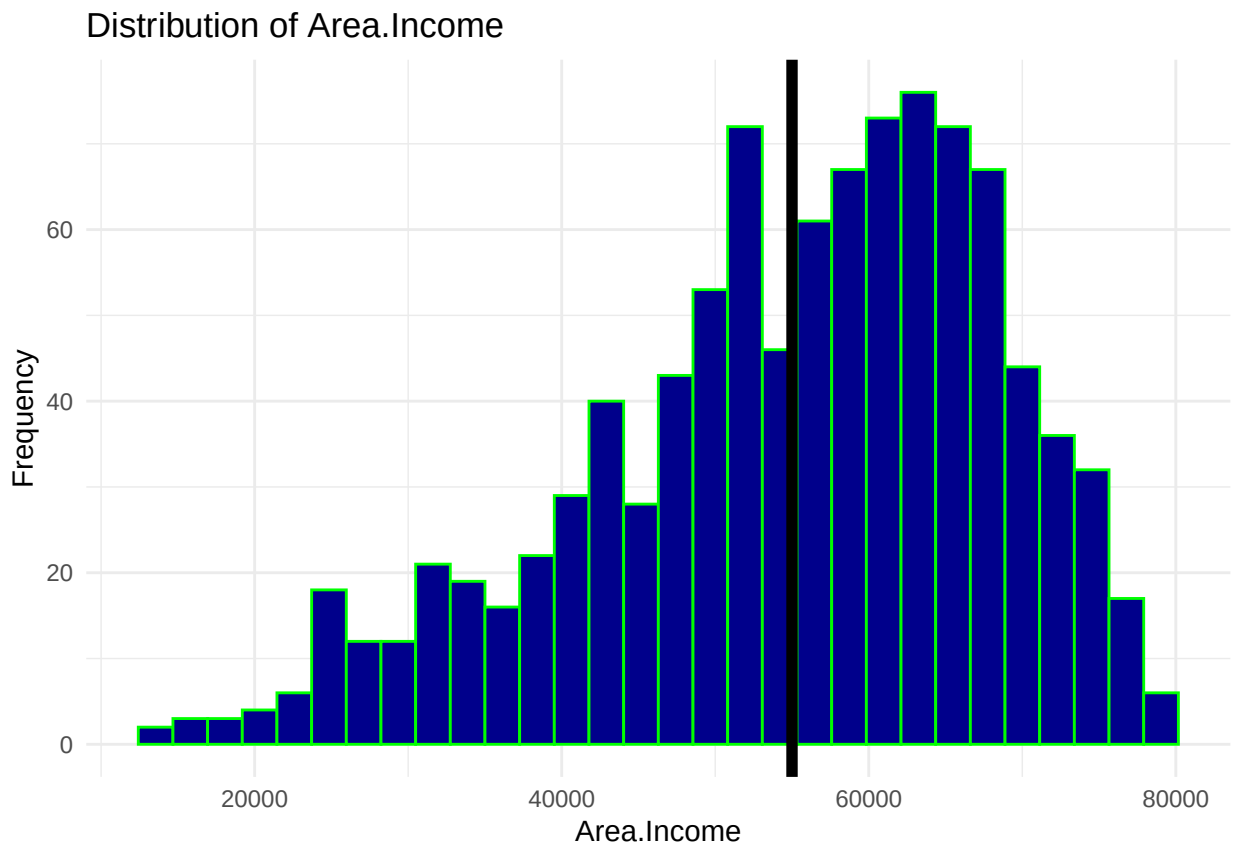

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

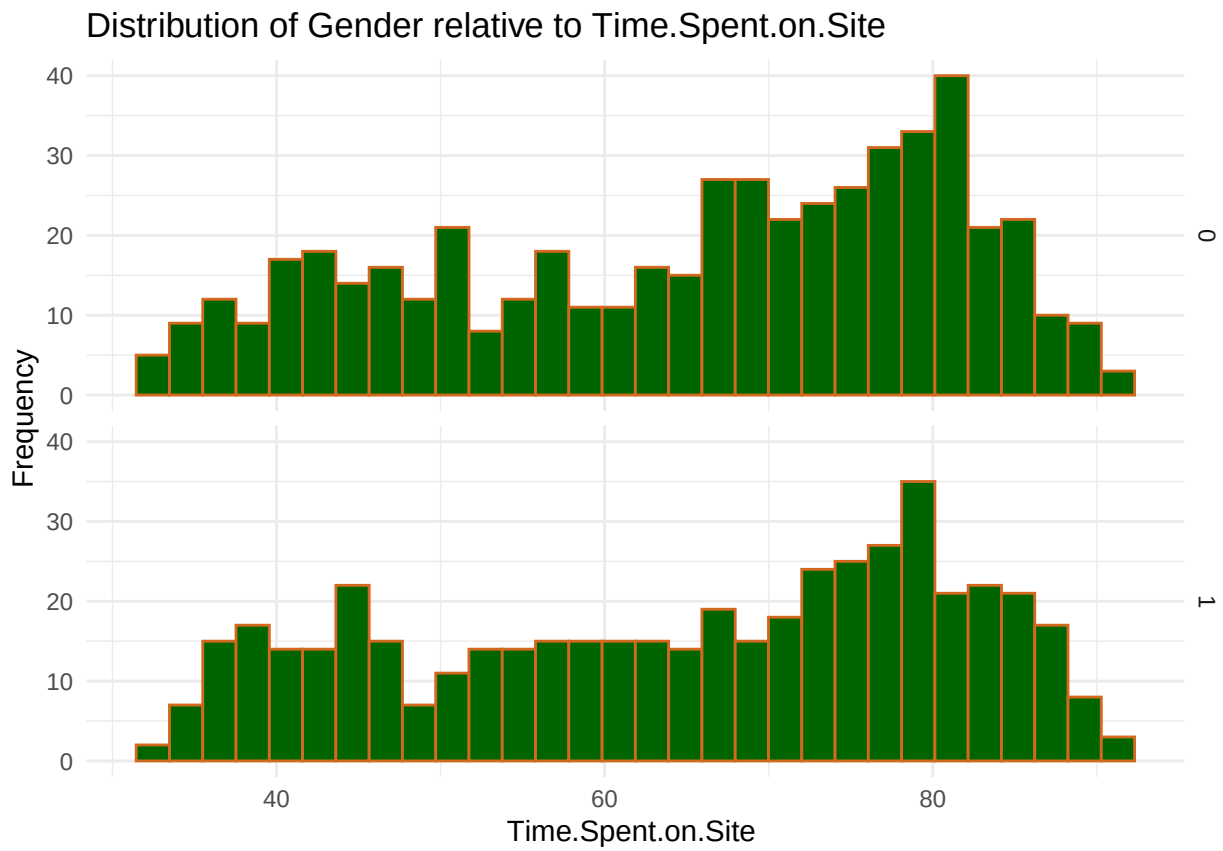



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

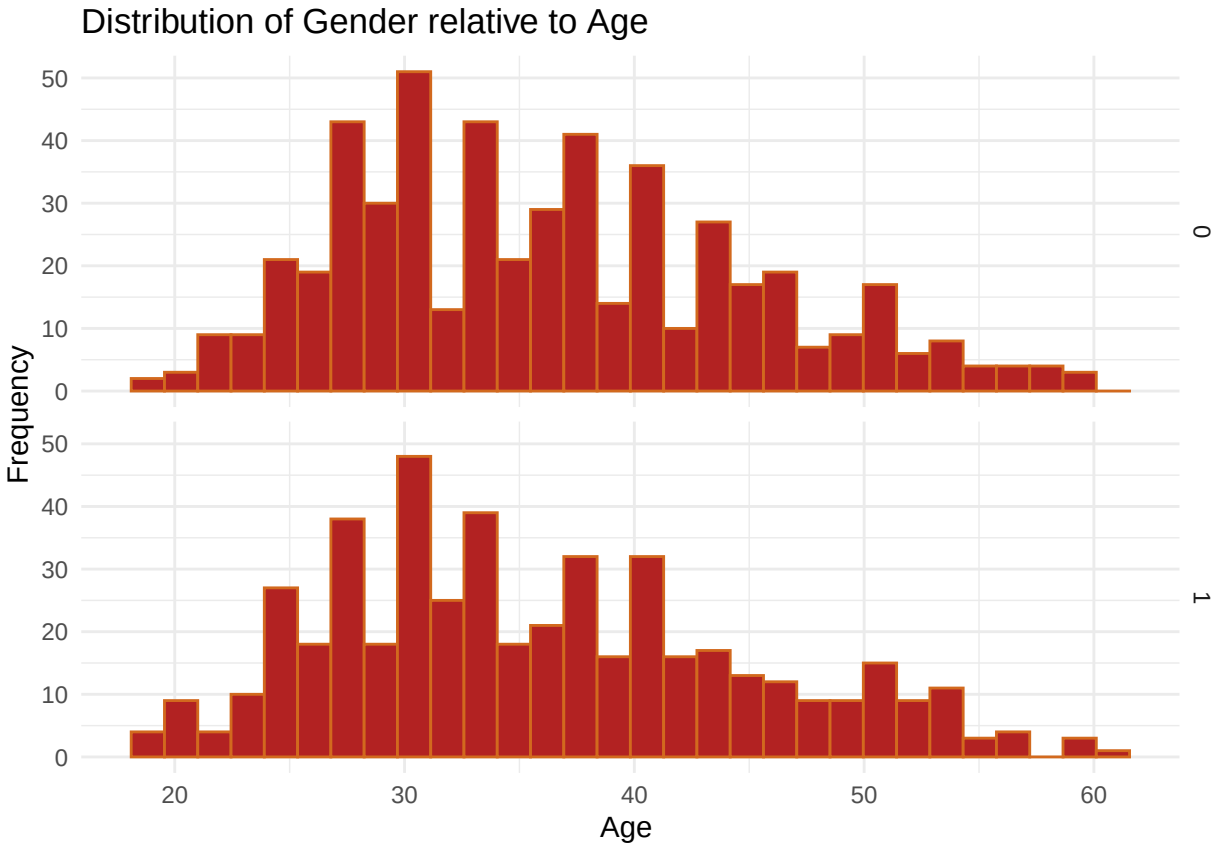


Bivariate analysis of a continuous variable with respect to a categorical variable

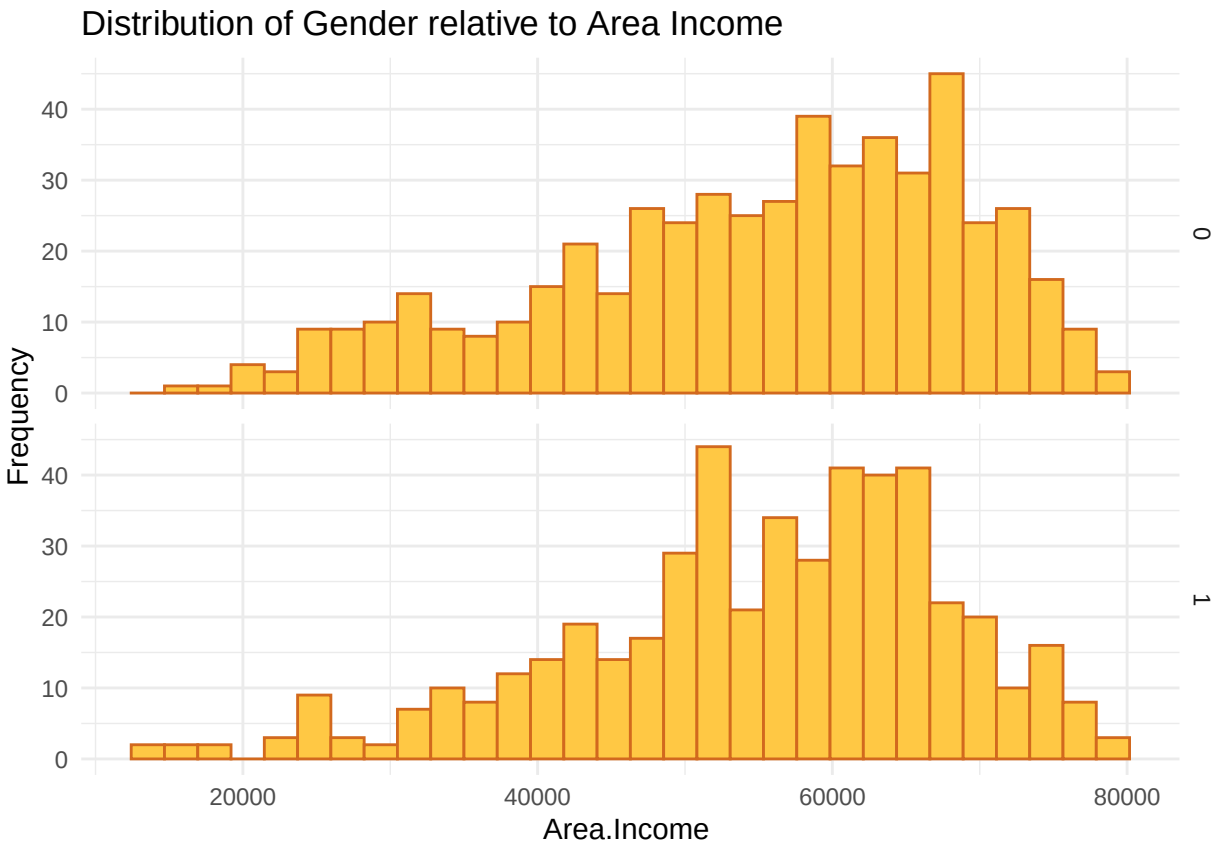
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

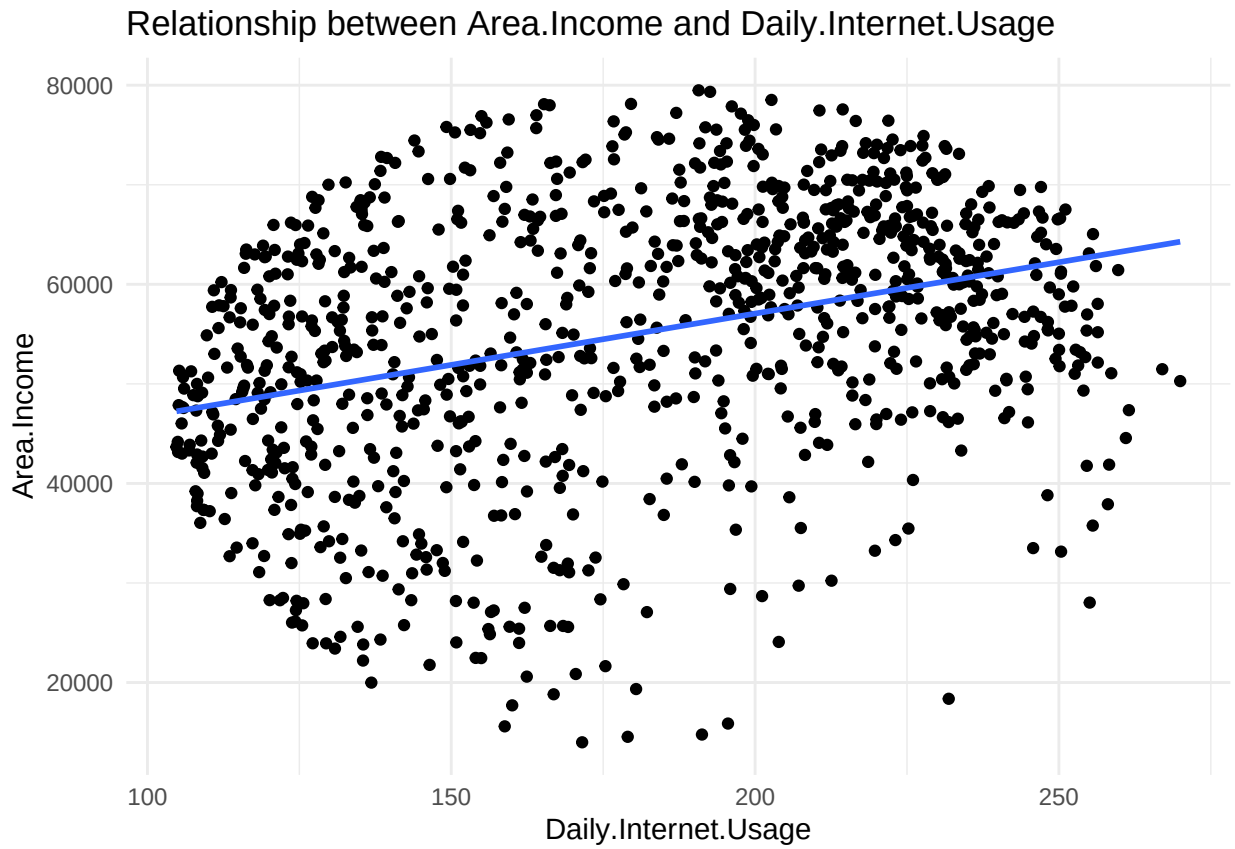



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

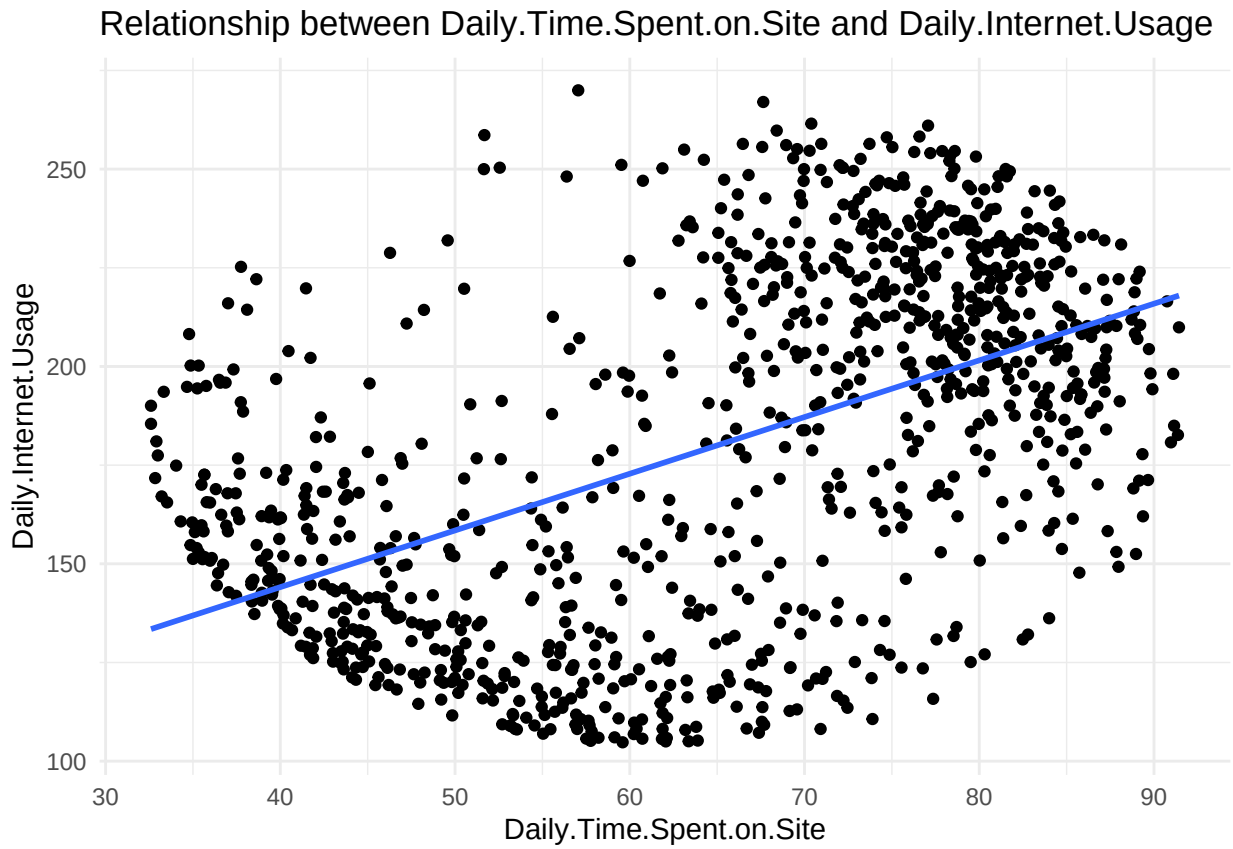


Bivariate analysis of a continuous variable with respect to another continuous variable

```
## `geom_smooth()` using formula 'y ~ x'
```

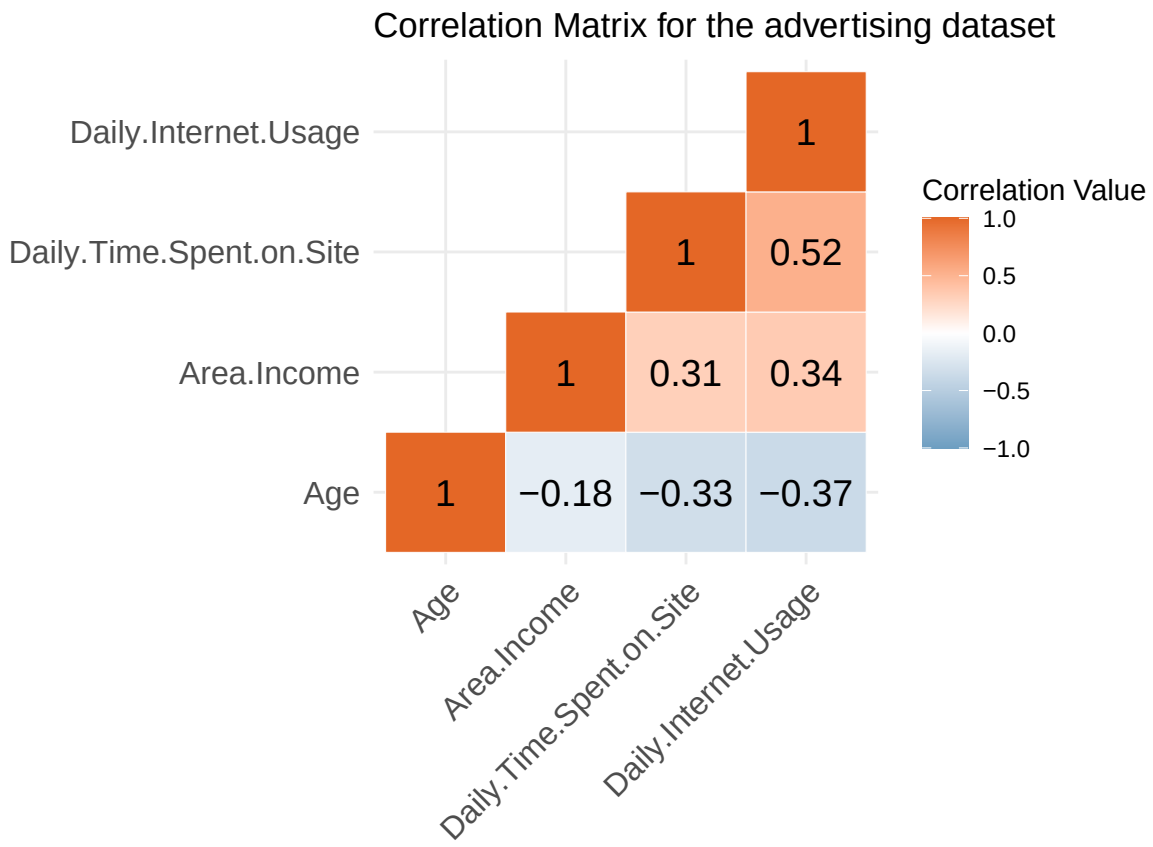



```
## `geom_smooth()` using formula 'y ~ x'
```



Correlation Matrix for the advertising dataset

```
## Warning: package 'ggcorrplot' was built under R version 4.0.5
```



ADVERTISMENT CLICK PREDICTION

```
## Warning: The `i` argument of ``[`()` can't be a matrix as of tibble 3.0.0.
## Convert to a vector.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## [1] 800   8
```

```
## [1] 200   8
```

```
##                                    Estimate           Std. Error
## (Intercept)              26.705967217933391566 3.00527054408254e+00
## Daily.Time.Spent.on.Site -0.195888728289965924 2.29526286859827e-02
## Age                       0.160387246318940024 2.69986786576503e-02
## Area.Income              -0.000133855039709939 2.03727500837067e-05
## Daily.Internet.Usage     -0.058865277529957773 6.98483182970521e-03
##                                    z value             Pr(>|z|)
## (Intercept)               8.88637705863727 6.31338418020445e-19
## Daily.Time.Spent.on.Site -8.53447903374989 1.40788362616520e-17
## Age                       5.94055910486171 2.84051620489459e-09
## Area.Income              -6.57029802849204 5.02146536843523e-11
## Daily.Internet.Usage     -8.42758694341280 3.52866087576811e-17
```

Plotting Predicted Probabilities

Now we will create a plot for each predictor. This can be very helpful for helping us understand the effect of each predictor on the probability of a 1 response on our dependent variable.

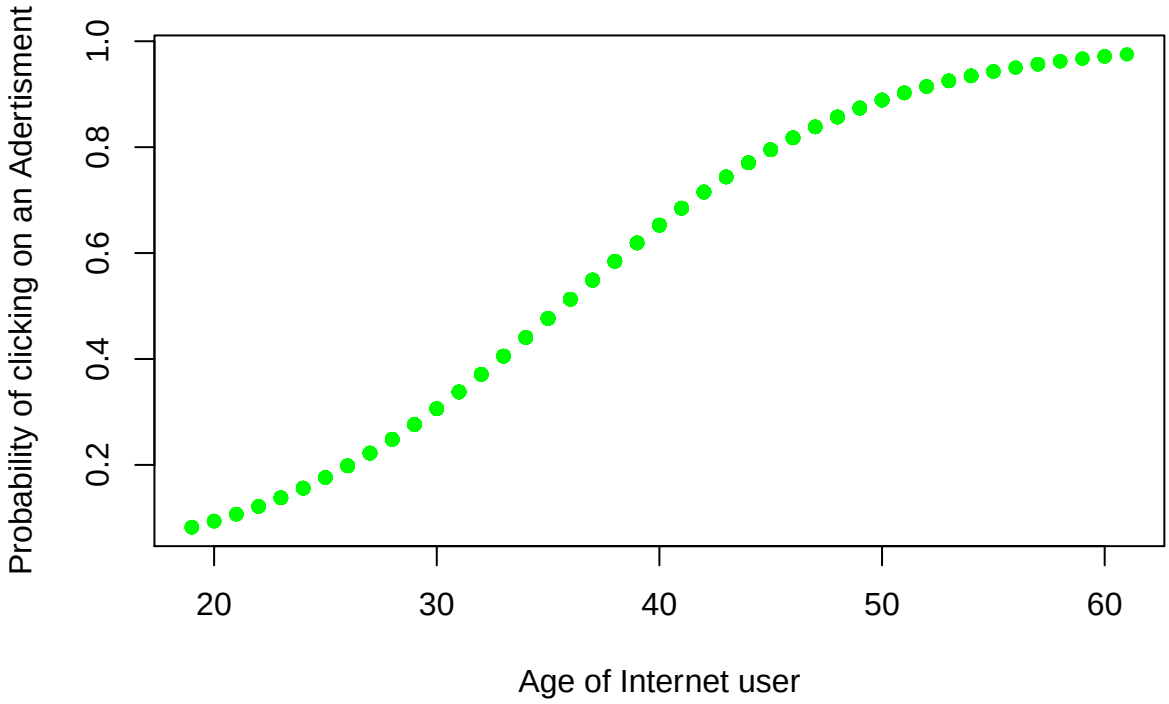
We wish to plot each predictor separately, so first we fit a separate model for each predictor. This isn't the only way to do it, but one that I find especially helpful for deciding which variables should be entered as predictors. The logistic function gives an s-shaped probability curve illustrated as follows:

```
##
## Call:
## glm(formula = knn_data$Clicked.on.Ad ~ knn_data$Age, family = binomial)
##
## Deviance Residuals:
##              Min                1Q            Median                3Q
## -2.2771393726158  -0.9080111115625  -0.0952709588871   0.9236362460542
##              Max
##  2.2350170484547
##
## Coefficients:
##                     Estimate      Std. Error   z value   Pr(>|z|)
## (Intercept)  -5.164948769293  0.375425650770 -13.75758 < 2.22e-16 ***
## knn_data$Age  0.144903106034  0.010474123573  13.83439 < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1386.294361120  on 999  degrees of freedom
## Residual deviance: 1112.280902457  on 998  degrees of freedom
## AIC: 1116.280902457
##
## Number of Fisher Scoring iterations: 4

## [1] 19 61

## [1] 1000

## [1] 1000
```



The model has produced a curve that indicates the probability of clicking on an Advertisement = 1 to Age. Clearly, the higher the Age, the more likely it is that one will click.

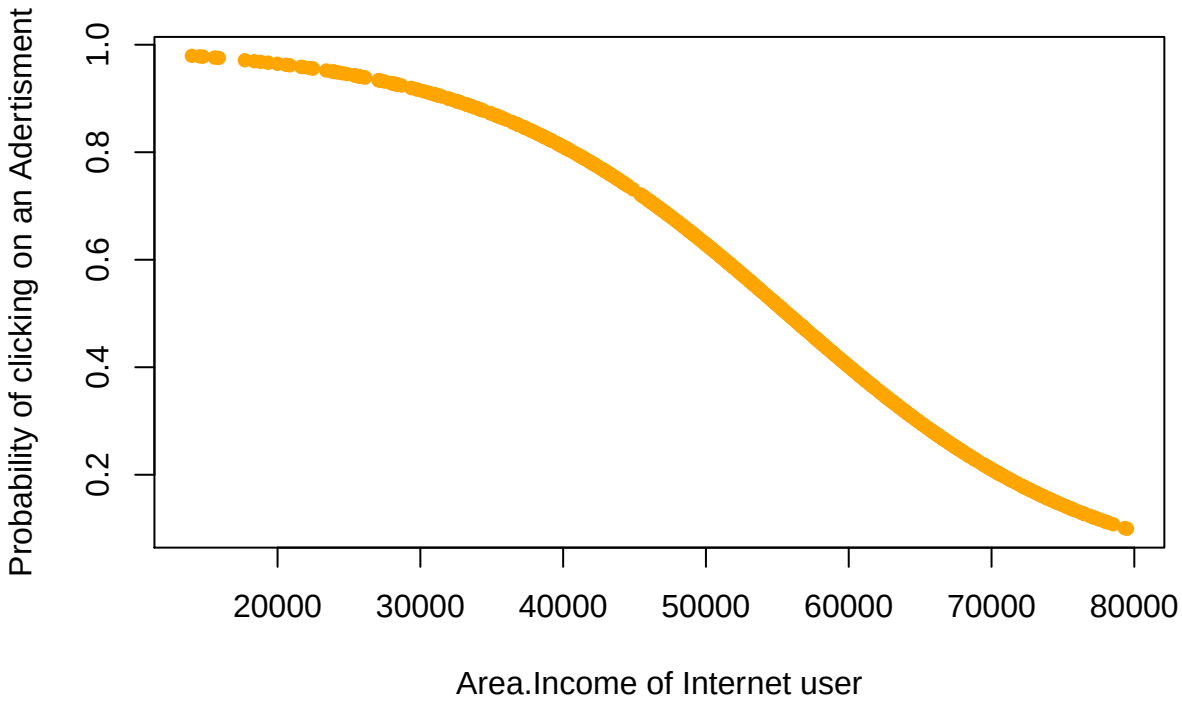
```
##
## Call:
## glm(formula = knn_data$Clicked.on.Ad ~ knn_data$Area.Income,
##     family = binomial)
##
## Deviance Residuals:
##             Min               1Q           Median               3Q
## -2.095568650429  -0.931274539515  -0.126851089507   0.940847563610
##             Max
##  2.110303785785
##
## Coefficients:
##                               Estimate         Std. Error   z value   Pr(>|z|)
## (Intercept)           5.15360176255e+00  3.97773412904e-01  12.95612 < 2.22e-16
## knn_data$Area.Income -9.25374572998e-05  6.91413280454e-06 -13.38381 < 2.22e-16
##
## (Intercept)          ***
## knn_data$Area.Income ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1386.294361120  on 999  degrees of freedom
```

```
## Residual deviance: 1128.914905226  on 998  degrees of freedom
## AIC: 1132.914905226
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] 13996.5 79484.8
```

```
## [1] 1000
```

```
## [1] 1000
```



Clearly, those who live in High income areas are unlikely to click on Advertisement.

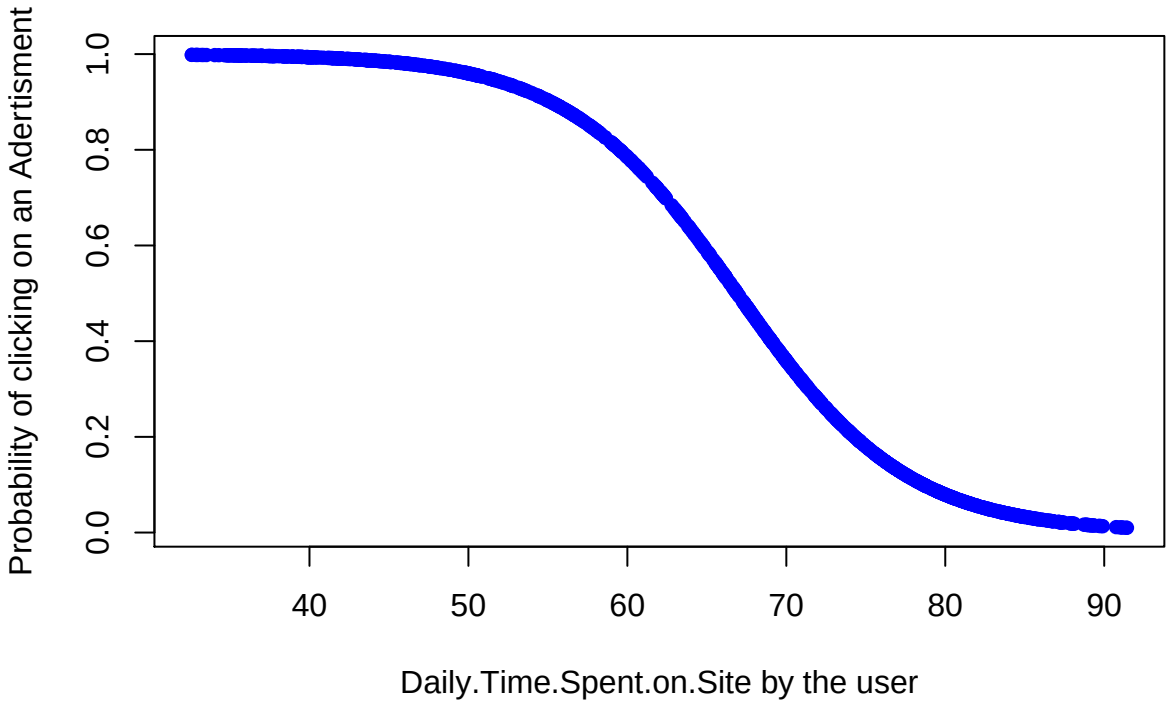
```
##
## Call:
## glm(formula = knn_data$Clicked.on.Ad ~ knn_data$Daily.Time.Spent.on.Site,
##     family = binomial)
##
## Deviance Residuals:
##              Min                1Q            Median                3Q
## -2.6579886896135  -0.5008089516518  -0.0424976458258   0.3296597118080
##              Max
##  3.0312867256044
##
## Coefficients:
##                                          Estimate      Std. Error   z value
## (Intercept)                          12.539951165562  0.795390449574  15.76578
## knn_data$Daily.Time.Spent.on.Site -0.187415350091  0.011502854316 -16.29294
##                                      Pr(>|z|)
```

```
## (Intercept)                        < 2.22e-16 ***
## knn_data$Daily.Time.Spent.on.Site < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1386.2943611199  on 999  degrees of freedom
## Residual deviance:  647.3146434587  on 998  degrees of freedom
## AIC: 651.3146434587
##
## Number of Fisher Scoring iterations: 6
```

```
## [1] 32.60 91.43
```

```
## [1] 1000
```

```
## [1] 1000
```



Clearly, those who spend more time on the internet are less likely to click on Advertisement.

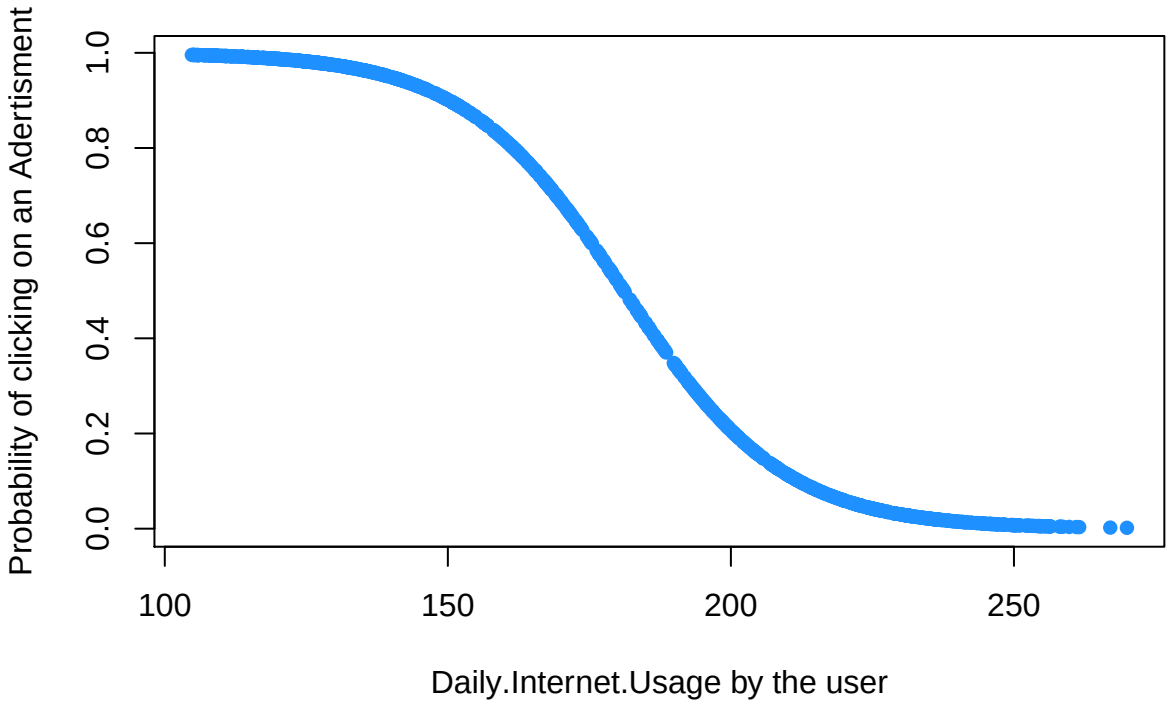
```
##
## Call:
## glm(formula = knn_data$Clicked.on.Ad ~ knn_data$Daily.Internet.Usage,
##     family = binomial)
##
## Deviance Residuals:
##             Min               1Q           Median               3Q
## -2.264037129037  -0.394769280082   0.013515998280   0.309926401427
```

```
##             Max
##  3.555868425564
##
## Coefficients:
##                                       Estimate        Std. Error   z value
## (Intercept)                    12.88034562405849  0.77616011942671  16.59496
## knn_data$Daily.Internet.Usage -0.07112404790199  0.00421091826939 -16.89039
##                                 Pr(>|z|)
## (Intercept)                   < 2.22e-16 ***
## knn_data$Daily.Internet.Usage < 2.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1386.2943611199  on 999  degrees of freedom
## Residual deviance:  570.5966465486  on 998  degrees of freedom
## AIC: 574.5966465486
##
## Number of Fisher Scoring iterations: 6
```

```
## [1] 104.78 269.96
```

```
## [1] 1000
```

```
## [1] 1000
```



Feature importance and Logistic Regression

```
## List of 1
##  $ axis.line:List of 6
##   ..$ colour       : chr "darkblue"
##   ..$ size         : num 0.5
##   ..$ linetype     : chr "solid"
##   ..$ lineend      : NULL
##   ..$ arrow        : logi FALSE
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_line" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
## Warning in countrycode(advertising_data$Country, origin = "country.name", : Some values were not mat
```

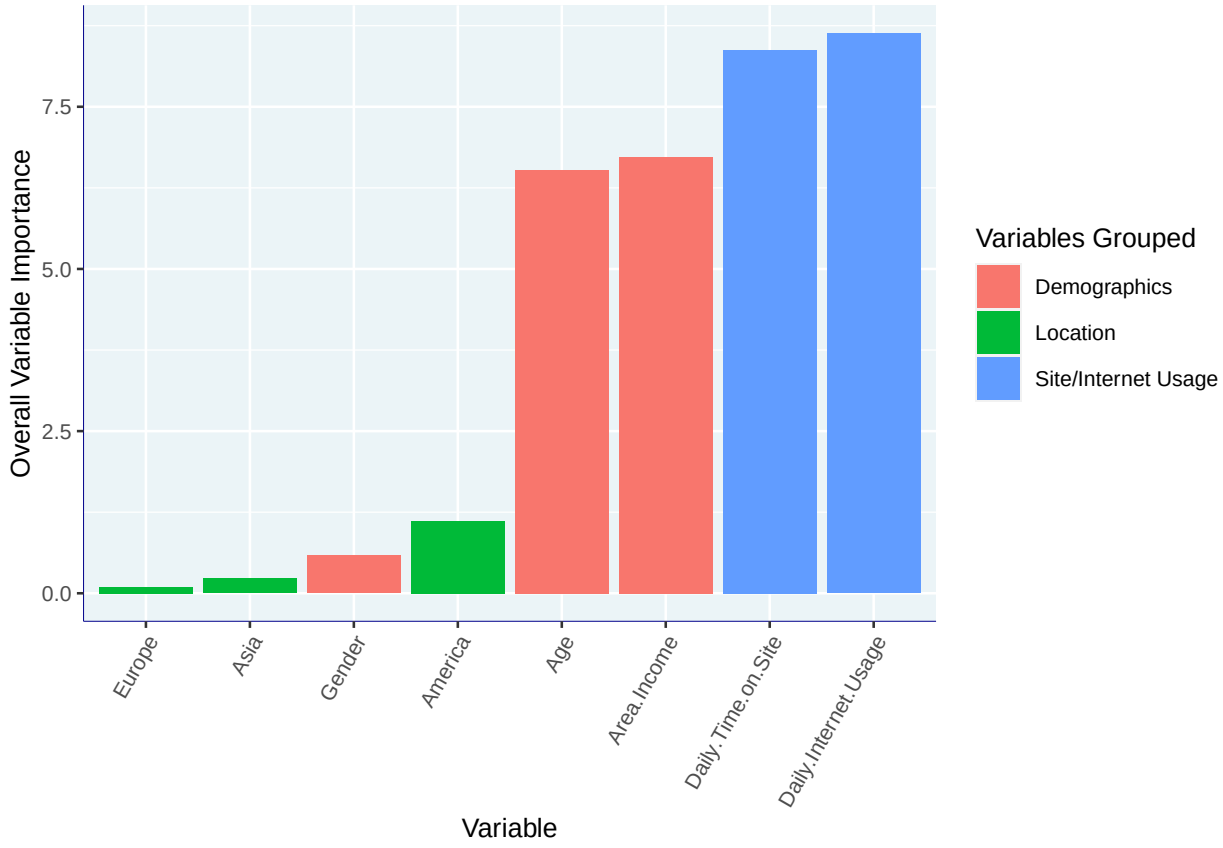
```
## [1] "Logistic Regression Confusion/Clarity Matrix)"
```

```
##    Class_predict
##       0   1
##   0 425   9
##   1  16 415
```

```
## [1] "Logistic Regression Accuracy:  97.1098265895954"
```

Table 3: Variable Importance:

	Overall	varnames	var_categ
Daily.Time.Spent.on.Site	8.379334095222884	Daily.Time.Spent.on.Site	Location
Age	6.527516885189965	Age	Location
Area.Income	6.724826176192777	Area.Income	Location
Daily.Internet.Usage	8.633131371671141	Daily.Internet.Usage	Location
Male	0.582364607079277	Male	Location
continent3	0.227296062920125	continent3	Location
continent4	0.098716143074804	continent4	Location
continent5	1.117696233696151	continent5	Location



Conclusion and recommendation

- Older persons are more likely to click on Advertisement.
- Those who use more internet are less likely to click on Advertisement.
- Those who live in High income areas are unlikely to click on Advertisement.
- Those who spend more time on the internet are less likely to click on Advertisement.