
Title: Cryptograph AdsS Prediction Author: Joseph Njuguna Date: 27/5/22

#1.Defining the question

#a) Specifying the question Identify individuals most likely to click on ads.

#b) Defining the metric for success Ability to identify individuals that click on ads.

#c) Understanding the context 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.

#d) Recording the experimental design -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. -From your insights provide a conclusion and recommendation.

#2. Reading the data

```
# choosing working directory that has uploaded file
getwd()
```

```
## [1] "C:/Users/jojo/Desktop/R"
```

```
#setwd("C:/Users/jojo/Downloads/R basics")
# using .csv to read dataset
# must have utils package installed.
adsop <- read.csv("advertising.csv", header= TRUE, sep= ",")
# view dataset
View(adsop)
```

#3. Checking the data

viewing first 5 rows of our dataset

```
head(adsop)
```

```
##   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
```

viewing last 5 rows of our dataset

```
tail(adsop)
```

```
##      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
```

checking data types

```
str(adsop)
```

```
## '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 standardi
##  $ 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" "
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...

# Our data types are numeric, integer, character.
```

shape of data

```
dim(adsop)
```

```
## [1] 1000   10

# Our dataset has 1000rows, 10 columns
```

descriptive statistical summary of our dataset

```
summary(adsop)
```

```
##  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
```

#4. Tidying the data

checking for duplicate records in our df

```
duplicates <- adsop[duplicated(adsop),]
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)

# No duplicate records in our dataset
```

missing values

list of columns and mising values

```
colSums(is.na(adsop))
```

```
## 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

# No missing values in our dataset.
```

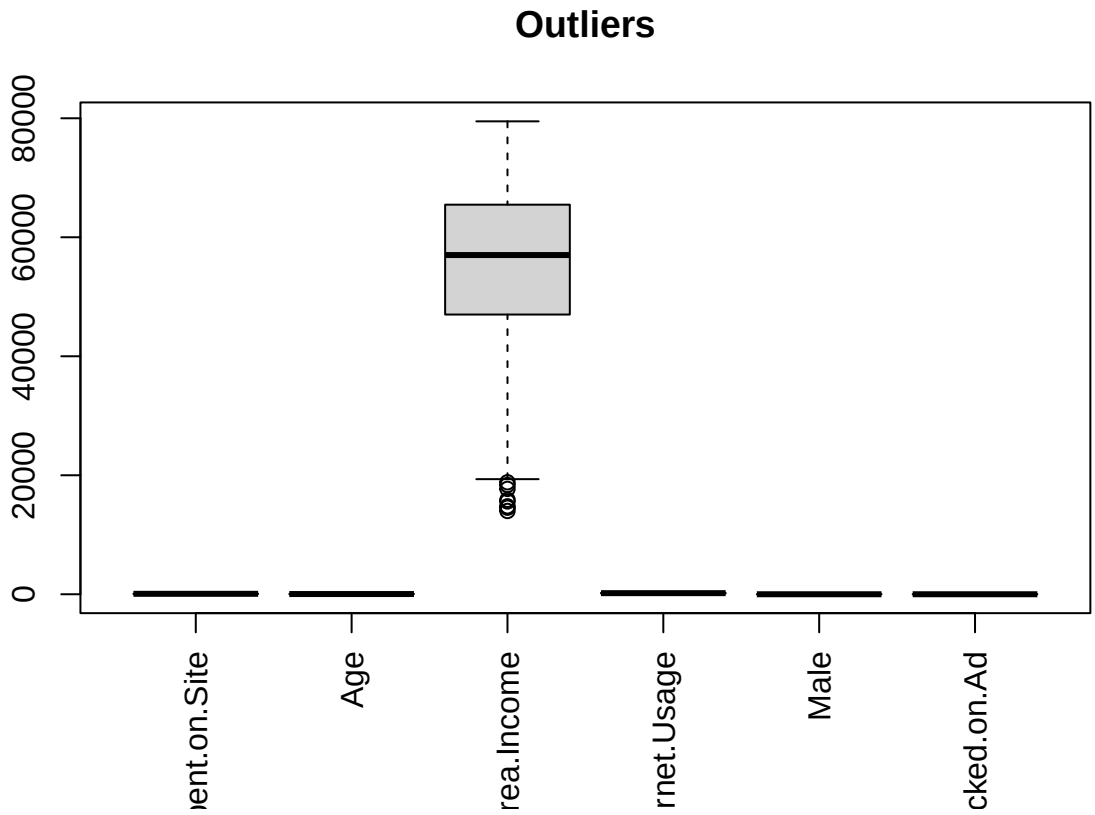
checking for outliers

listing numerical columns as we can only get outliers for numerical columns

```
numerical <- list(adsop$Daily.Time.Spent.on.Site,adsop$Age,
adsop$Area.Income,adsop$Daily.Internet.Usage,adsop$Male, adsop$Clicked.on.Ad)
```

creating boxplots

```
boxplot(numerical, names=c('Daily.Time.Spent.on.Site', 'Age', 'Area.Income', 'Daily.Internet.Usage', 'Ma
```



Outliers only exist in our area.income column # It is not necessary to remove them.

#5. Data Anlaysis # Univariate Analysis # Measures of central tendecy

Mean

```
colMeans(adsop[sapply(adsop, is.numeric)])
```

```
## Daily.Time.Spent.on.Site                      Age              Area.Income
##                  65.0002                 36.0090                 55000.0001
##     Daily.Internet.Usage                     Male            Clicked.on.Ad
##                 180.0001                  0.4810                  0.5000
```

```
# The mean age of respondents is 36 years.
# Mean area income is $55,000.
# Mean time spent on site daily is 65 minutes.
```

Median

daily time spent on site

```
median(adsop$Daily.Time.Spent.on.Site)
```

```
## [1] 68.215
```

age

```
median(adsop$Age)
```

```
## [1] 35
```

area income

```
median(adsop$Area.Income)
```

```
## [1] 57012.3
```

daily internet usage

```
median(adsop$Daily.Internet.Usage)
```

```
## [1] 183.13
```

Measures of dispersion

Variance

daily time spent on site

```
var(adsop$Daily.Time.Spent.on.Site)
```

```
## [1] 251.3371
```

age

```
var(adsop$Age)
```

```
## [1] 77.18611
```

area income

```
var(adsop$Area.Income)
```

```
## [1] 179952406
```

daily internet usage

```
var(adsop$Daily.Internet.Usage)
```

```
## [1] 1927.415
```

Standard deviation

daily time spent on site, age, area income, internet usage, male, clicked on ad

```
sd(adsop$Daily.Time.Spent.on.Site)
```

```
## [1] 15.85361
```

```
sd(adsop$Age)
```

```
## [1] 8.785562
```

```
sd(adsop$Area.Income)
```

```
## [1] 13414.63
```

```
sd(adsop$Daily.Internet.Usage)
```

```
## [1] 43.90234
```

```
sd(adsop$Male)
```

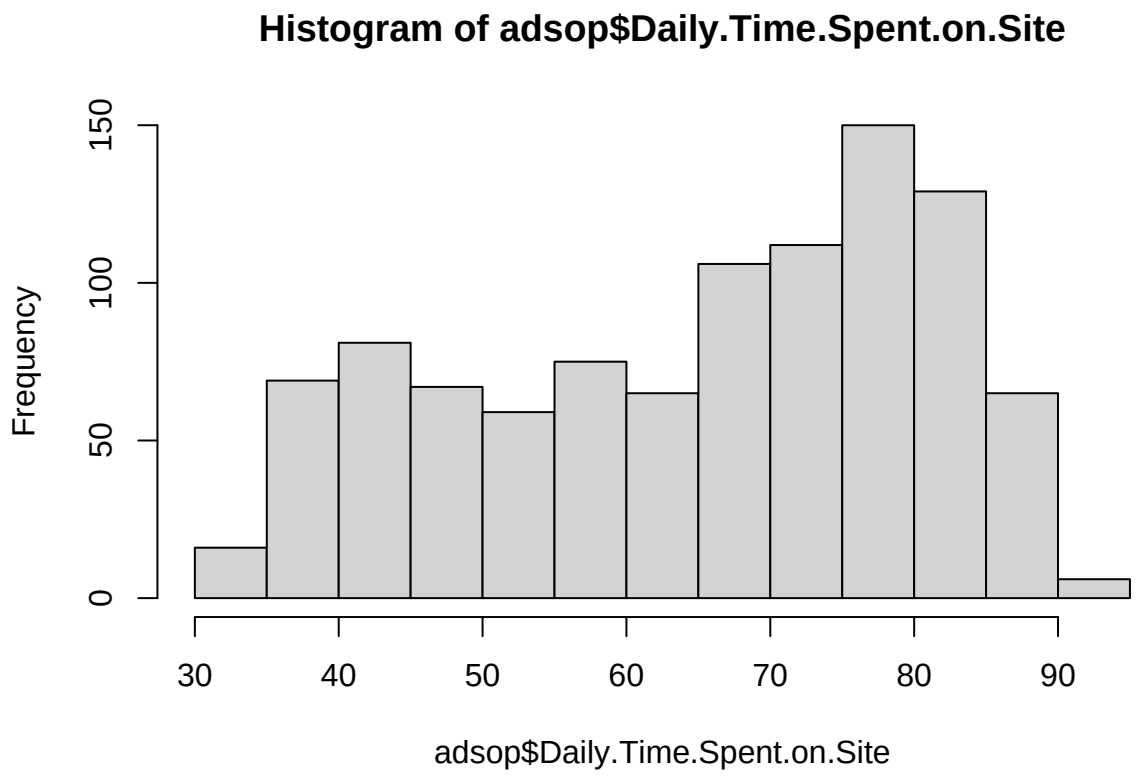
```
## [1] 0.4998889
```

```
sd(adsop$Clicked.on.Ad)
```

```
## [1] 0.5002502
```

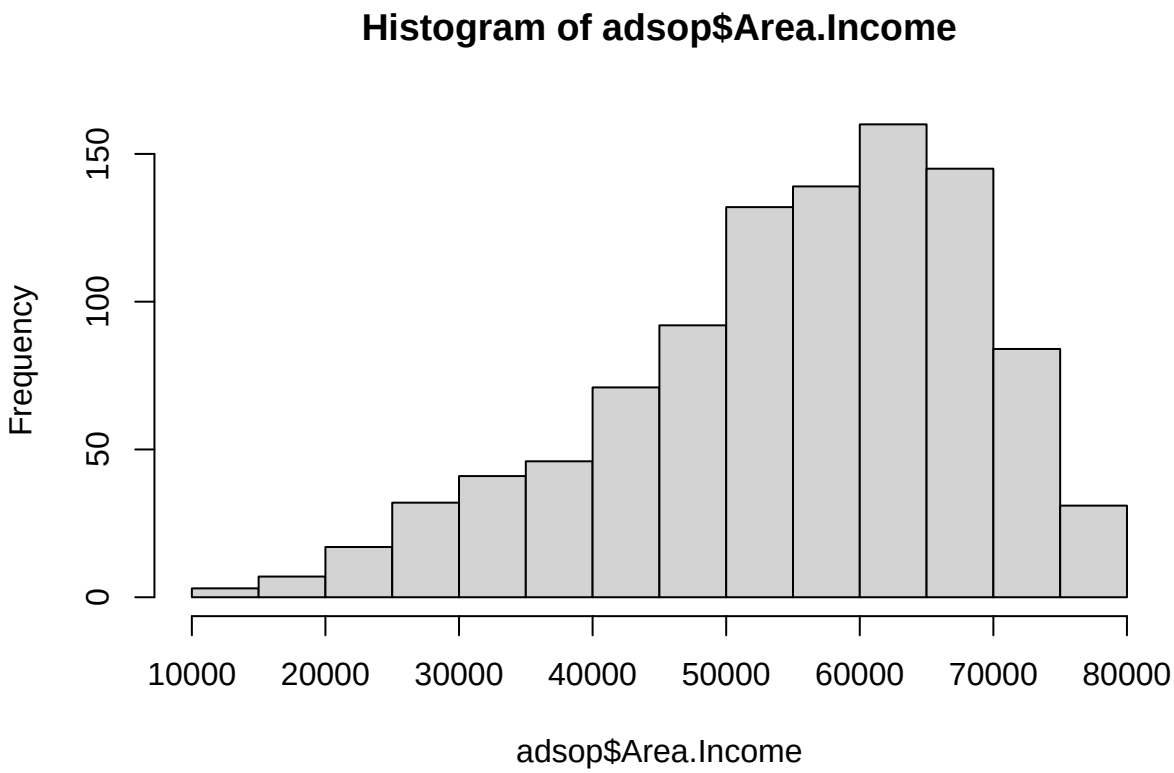
histogram - time spent on site

```
hist(adsop$Daily.Time.Spent.on.Site)
```



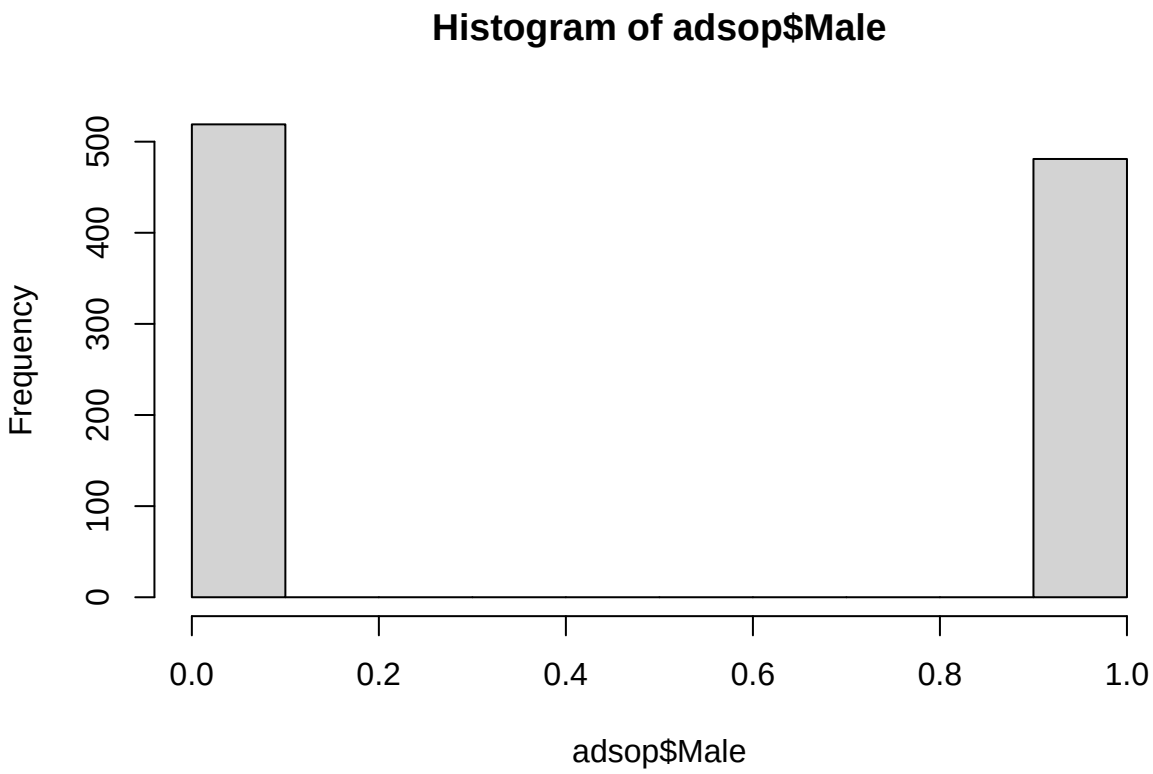
80 minutes is the most frequent time spent by users on site # hist - Area income

```
hist(adsop$Area.Income)
```



Highest area income revenue is $60,000 # histogram on gender distribution

```
hist(adsop$Male)
```

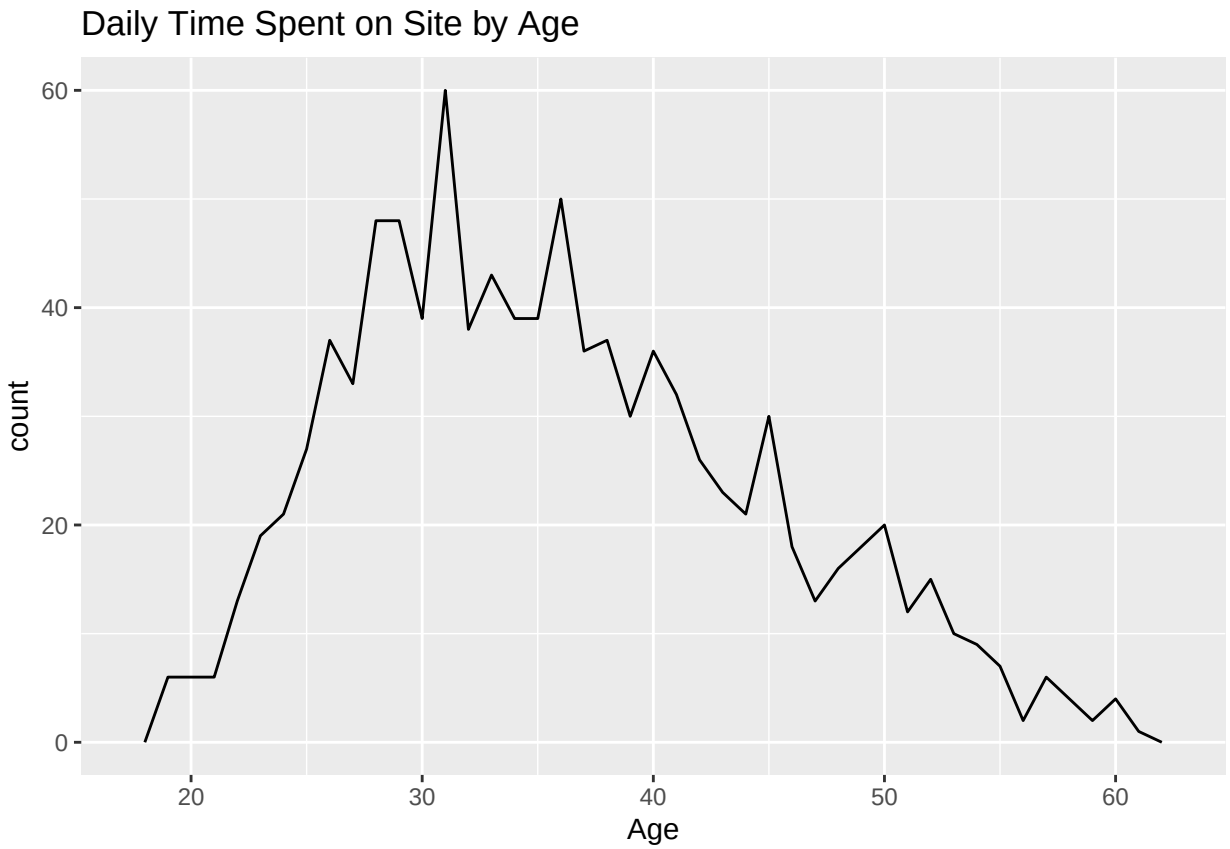


Number of female respondents were slightly higher than male respondents.

Bivariate Analysis

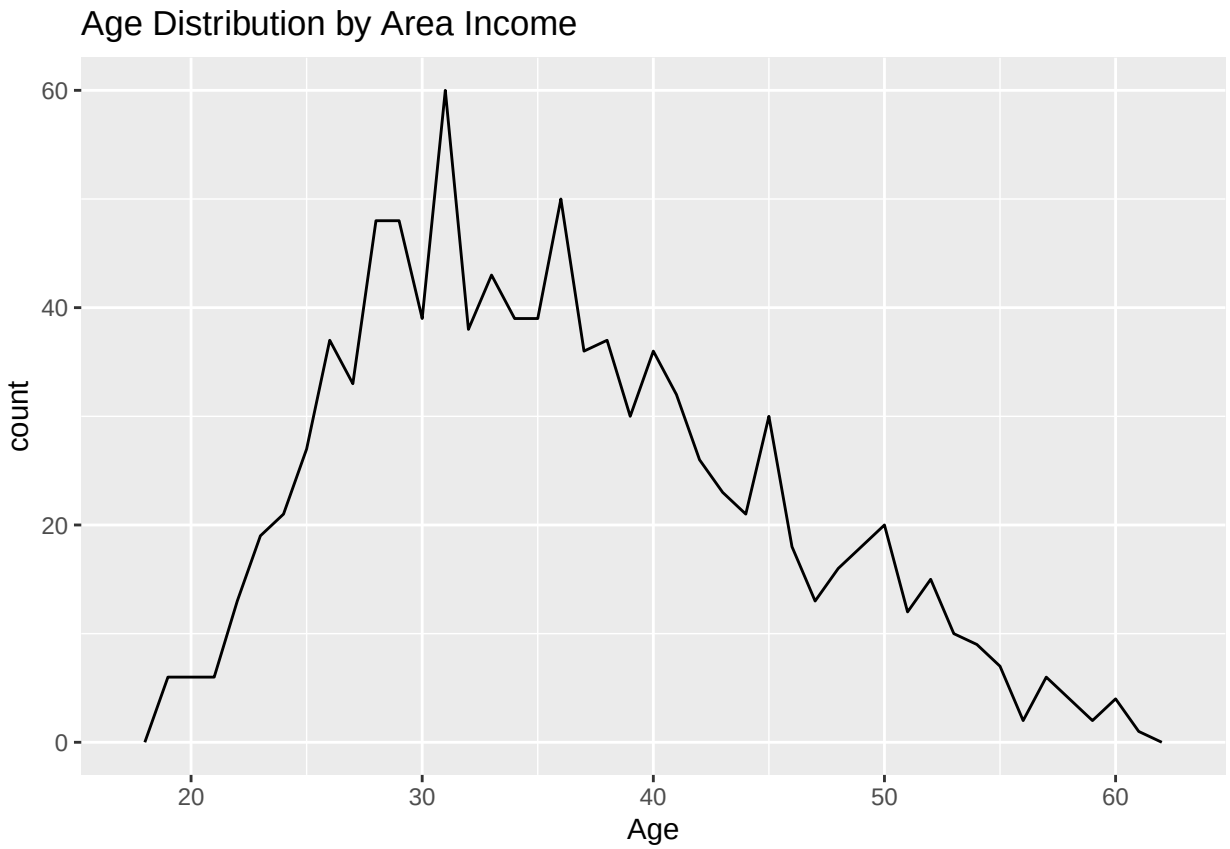
Numerical-Numerical variables

```
#ViZ of daily time spent on site by age
# ---
#
library(ggplot2)
ggplot(adsop, aes(Age, colour = Daily.Time.Spent.on.Site)) +
geom_freqpoly(binwidth = 1) + labs(title="Daily Time Spent on Site by Age")
```



We can observe that the highest amount of daily time spent on site increases steadily from age group 23-33 and decreases after that.

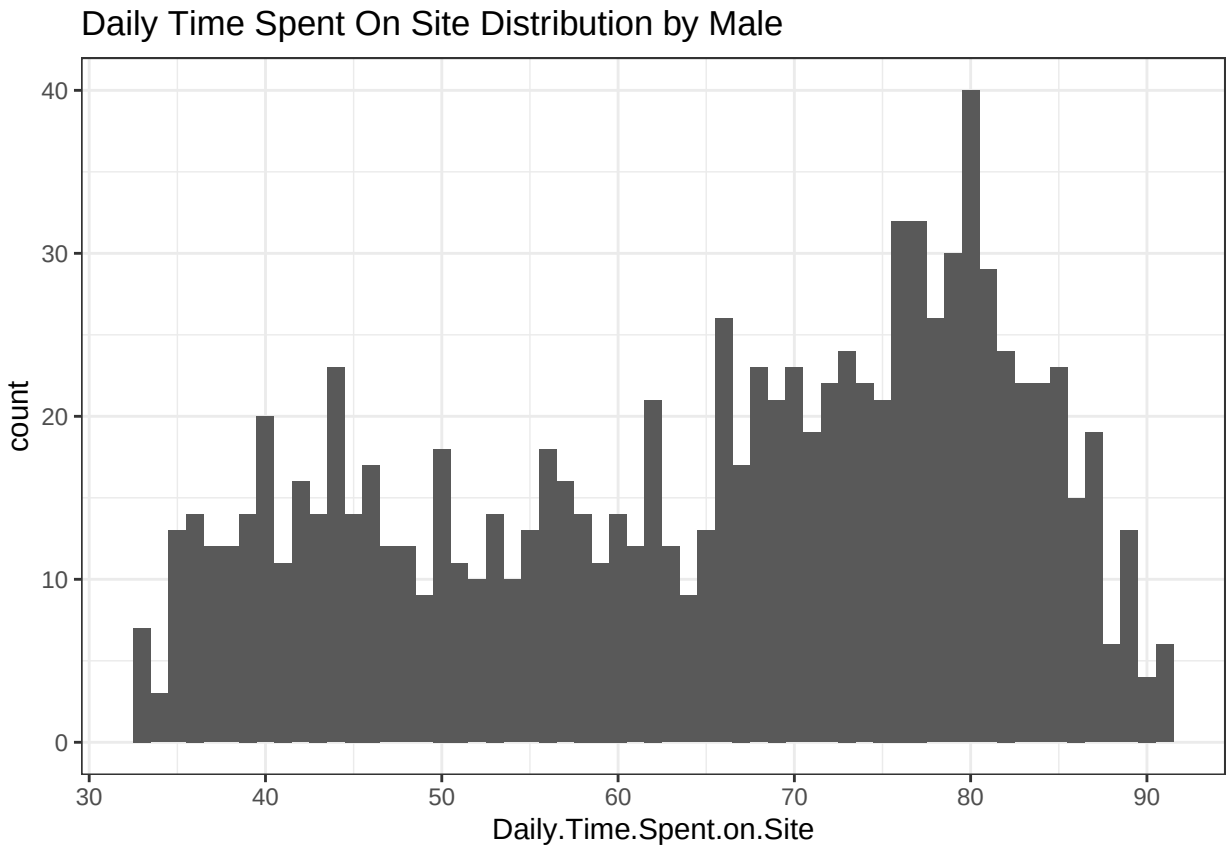
```
ggplot(adsop, aes(Age, colour = Area.Income)) +
geom_freqpoly(binwidth = 1) + labs(title="Age Distribution by Area Income")
```



We observe that area income was highest around the 30-33 years age group.

Numerical-Categorical

```
c <- ggplot(adsop, aes(x=Daily.Time.Spent.on.Site, fill=Male, color=Male)) +
geom_histogram(binwidth = 1) + labs(title="Daily Time Spent On Site Distribution by Male")
c + theme_bw()
```



#6. Conclusion - Average of 60 min is spent on the site per day.

#7. Recommendations - Predictive modelling could be conducted to predict certain outcomes.

#8. Follow up questions

#a) Did we have right data? Yes. #b) Do we need other data to answer our question? No. #c) Did we have the right question? Yes.

```
#ffghgff
```