Cryptoghraphy-EDA

Moureen

2022-08-04

Loading libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     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::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
library(moments)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

Loading the data

```
data <- read.csv("http://bit.ly/IPAdvertisingData")
head(data)
```

```
##   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 our data

```
# Structure of the data
str(data)
```

```
## '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 ...
```

```
# It is comprised of 1000 observations and 10 variables with numeric,character and integer data.
```

```
# Checking for data summary for our numeric and integers
summary(data)
```

```
##  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
```

Cleaning the Data

```
# Checking for duplicates
anyDuplicated(data)
```

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

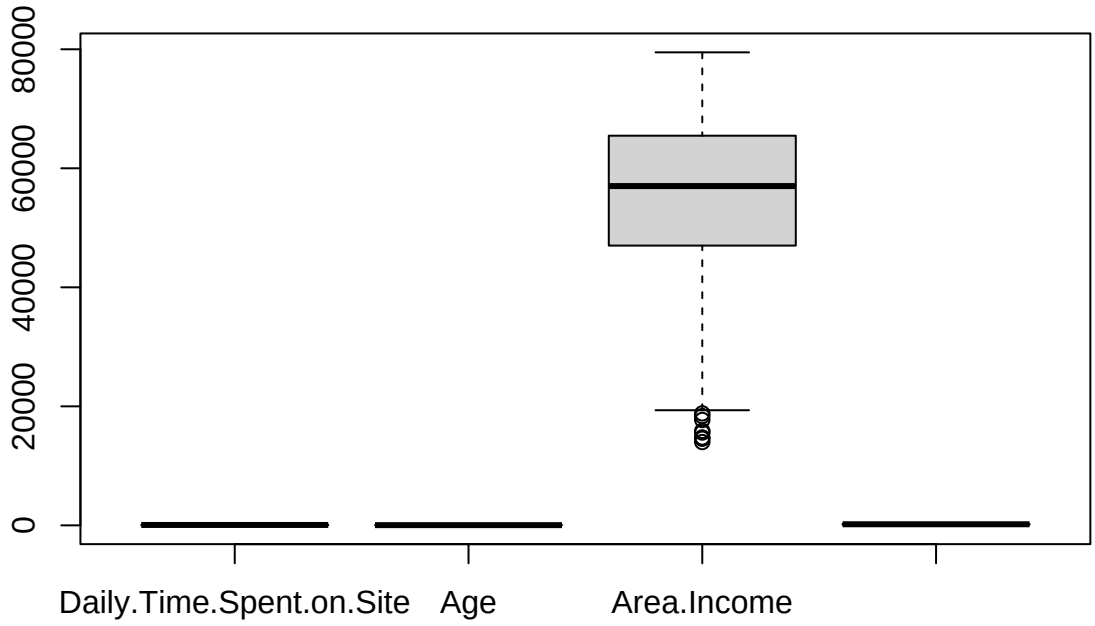
```
# There were no duplicates
```

```
# Checking for missing values
colSums(is.na(data))
```

```
## 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 were not missing values.
```

```
# Checking for outliers
# Put our numeric columns under one subset then proceed to draw our boxplots
df <- subset(data, select = -c(Ad.Topic.Line,City,  Male,   Country,    Timestamp,  Clicked.on.Ad))
boxplot(df)
```




```
# There are a few outliers on the Area.Income.
# Given that it is income, the figures will vary because we don't know if they were standardised to a c
```

Dealing with data types

```
# Changing timestamp to datetime
data$Timestamp <- as.POSIXct(data$Timestamp, "%Y-%m-%d %H:%M:%S",tz = "GMT")
head(data)
```

```
##   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
```

```
# Split the time and date
data$date = format(data$Timestamp, "%y/%m/%d")
data$time = format(data$Timestamp, "%H:%M:%S")
data$date <- as.Date(data$date)
```

```
# Drop column timestamp
df1 = subset(data, select = -c(Timestamp))
```

```
# Changing to factors
data$Male <- as.factor(data$Male)
data$Clicked.on.Ad <- as.factor(data$Clicked.on.Ad)
```

Exploratory Analysis

```
# Selecting our nmeric columns under one subset
num <- subset(df1, select = -c(Ad.Topic.Line,City,  Male,   Country,    date,   Clicked.on.Ad, time))
```

```
# Getting the summary  to observe the measures of our central tendencies.
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
```

Observations

The mean age of people visiting the site is 36, max age is 61 and min age is 19 which makes sense since the range between 61 and 19 are the people most active online.

The mean daily internet usage on the website is 180 and a median level at 183.1

The minimum amount of time spent on the blog is 32.60 and maximum is 91.43 with a mean at 65 and median at 68

The maximum income of individuals is 79485 and a min income of 13996 and the mean being 55000

```
# Measures of dispersion
kurtosis(num)
```

```
## Daily.Time.Spent.on.Site                      Age              Area.Income
##                 1.903942                 2.595482                 2.894694
##     Daily.Internet.Usage
##                 1.727701
```

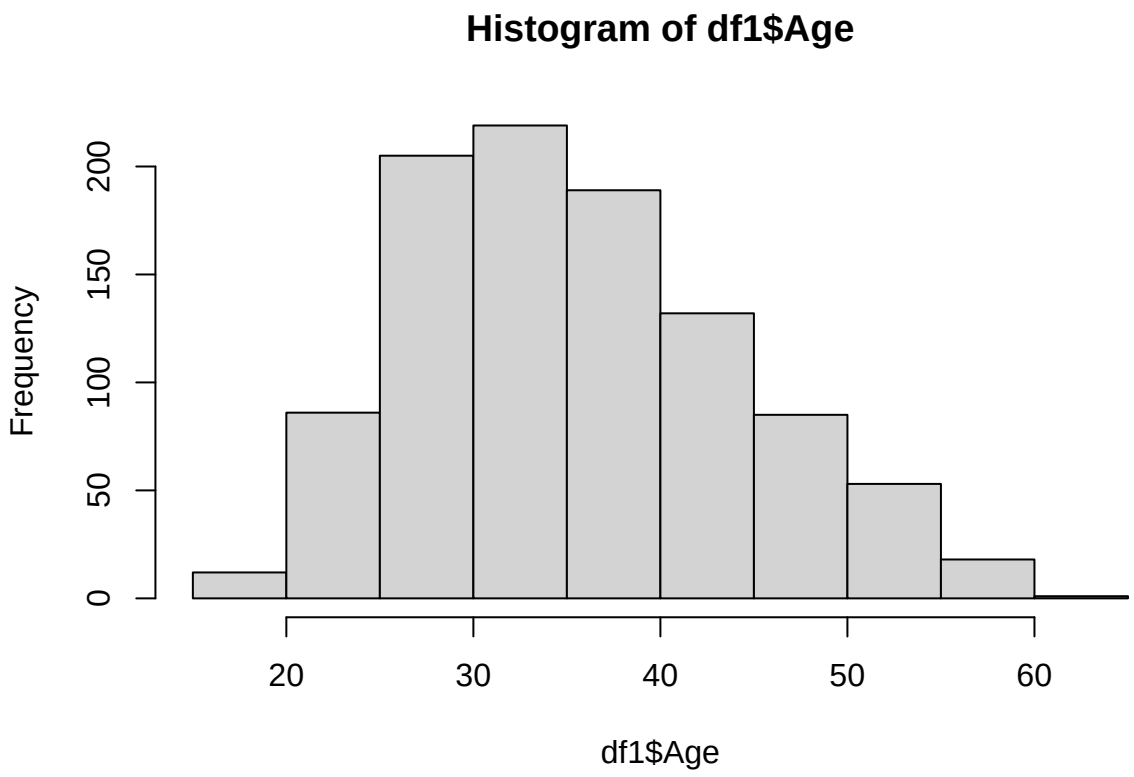
```
skewness(num)
```

```
## Daily.Time.Spent.on.Site                      Age              Area.Income
##              -0.37120261              0.47842268              -0.64939670
##     Daily.Internet.Usage
##              -0.03348703
```

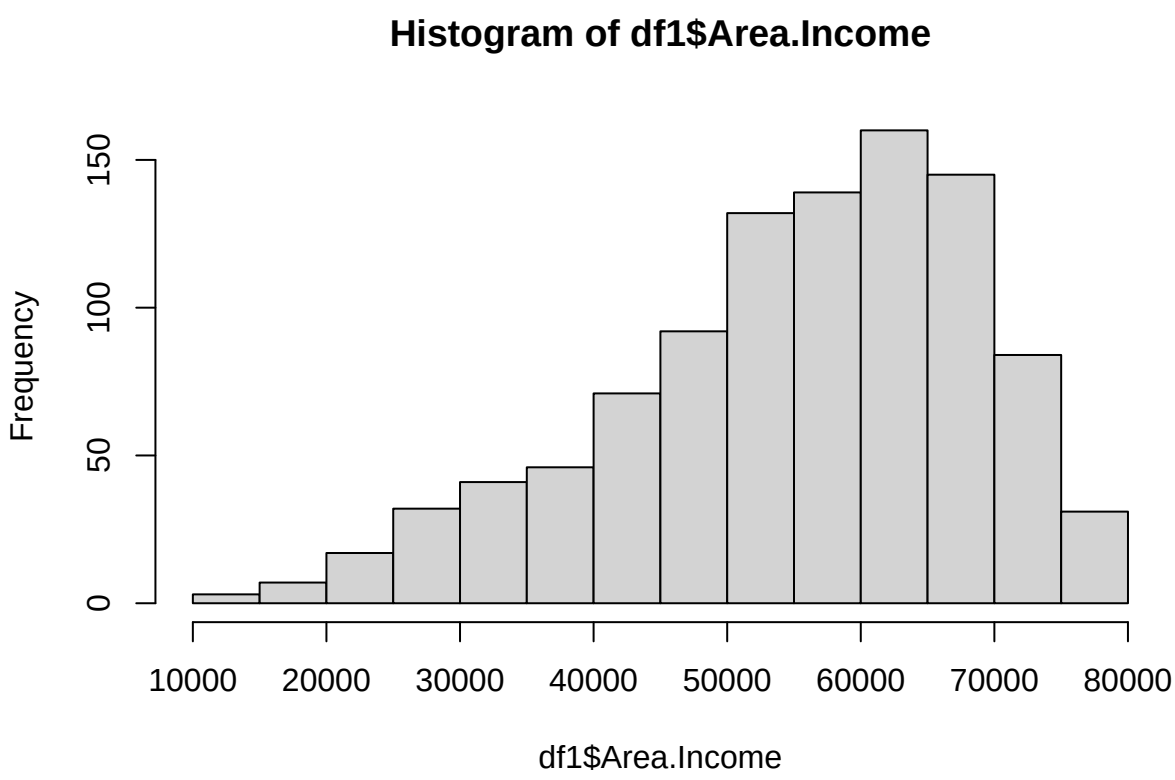
```
# The variance of the nmeric data
var(num)
```

```
##                           Daily.Time.Spent.on.Site          Age  Area.Income
## Daily.Time.Spent.on.Site                251.33709    -46.17415     66130.81
## Age                                     -46.17415     77.18611    -21520.93
## Area.Income                           66130.81091 -21520.92580 179952405.95
## Daily.Internet.Usage                    360.99188   -141.63482    198762.53
##                           Daily.Internet.Usage
## Daily.Time.Spent.on.Site             360.9919
## Age                                 -141.6348
## Area.Income                       198762.5315
## Daily.Internet.Usage                1927.4154
```

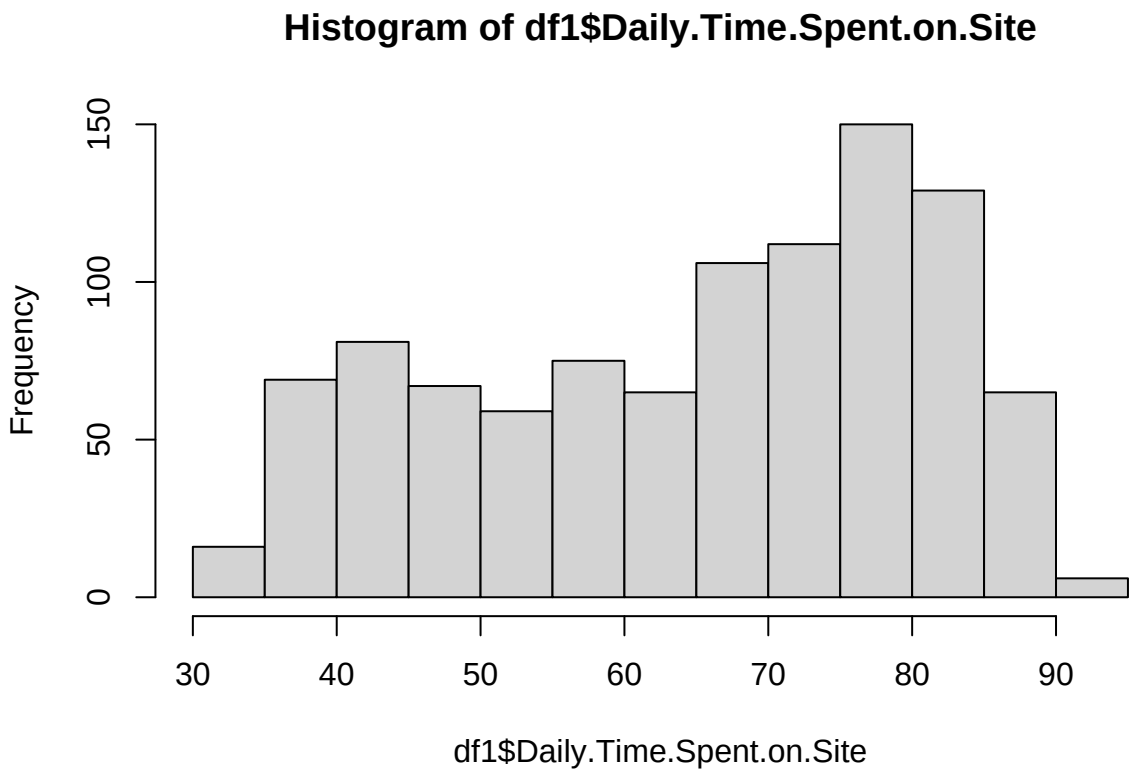
```
# Histogram to visualise distribution
hist(df1$Age)
```



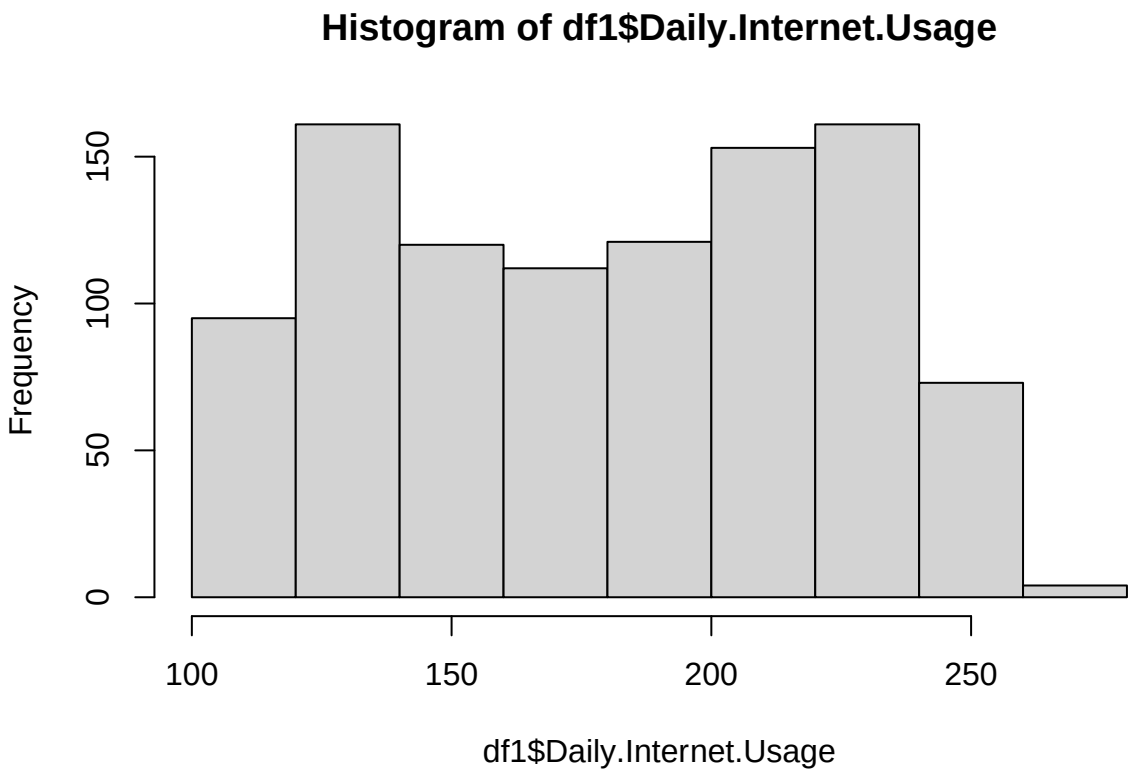

```
hist(df1$Area.Income)
```




```
hist(df1$Daily.Time.Spent.on.Site)
```




```
hist(df1$Daily.Internet.Usage)
```



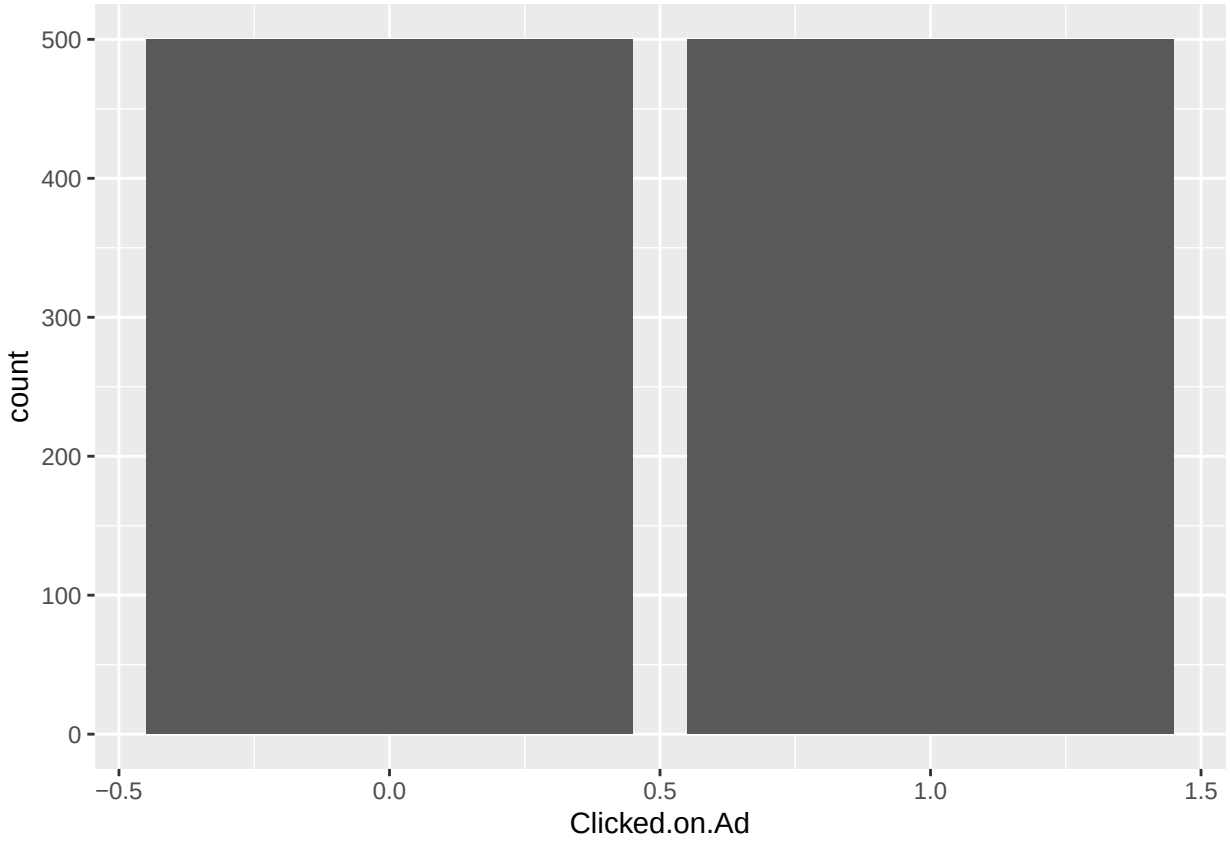
Observations

#Time spent on site: There range of spending is between 65 and 85 time on the site.

#Age: Most people who visit the blog are between 25 and 40 years, data is skewed to the right of the mean. Therefore its a positive skew

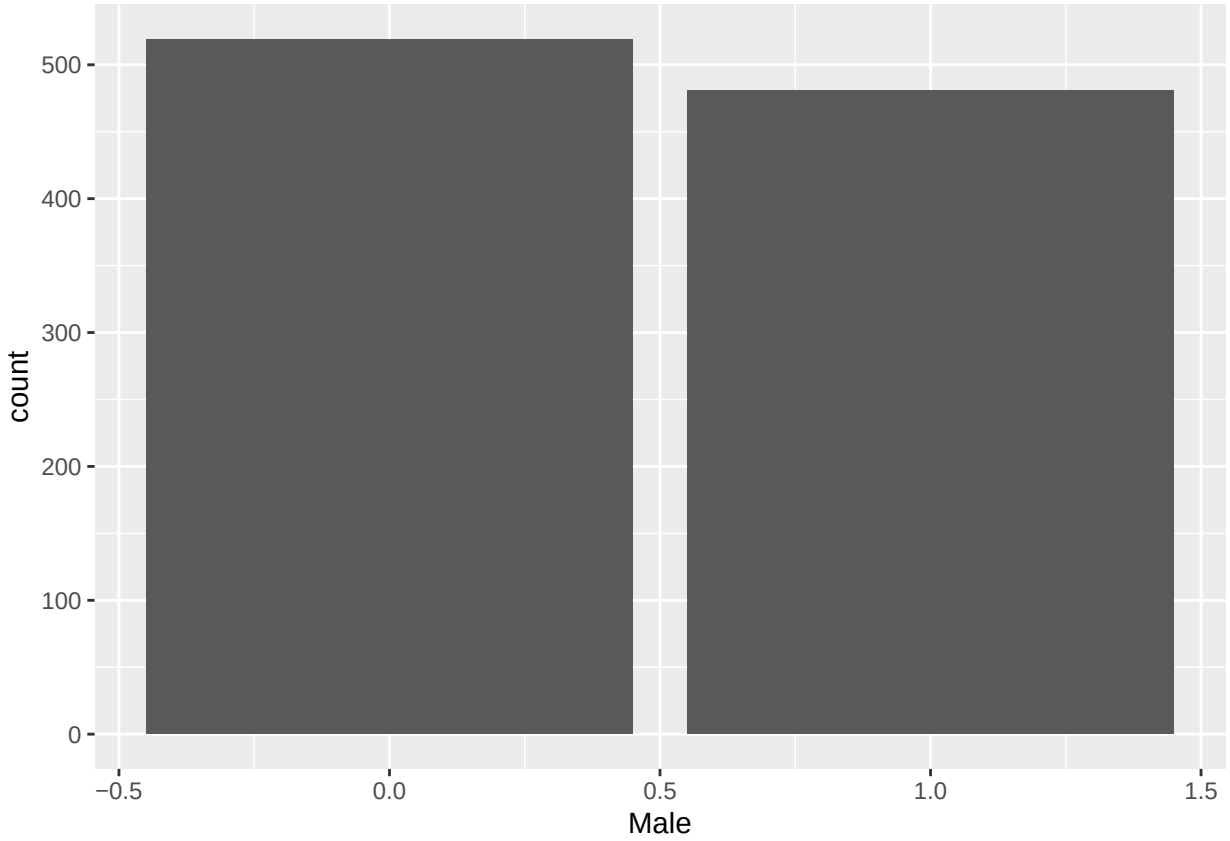
#Income: Data on income is mostly skewed to the right of the 55,00 mean, therefore a positive skew

```
# Distribution of non-numeric data
# clicked on ad
ggplot(data = df1) +
  geom_bar(mapping = aes(x = Clicked.on.Ad))
```




```
# This shows a balance between those who clicked on ads and those who did not
```

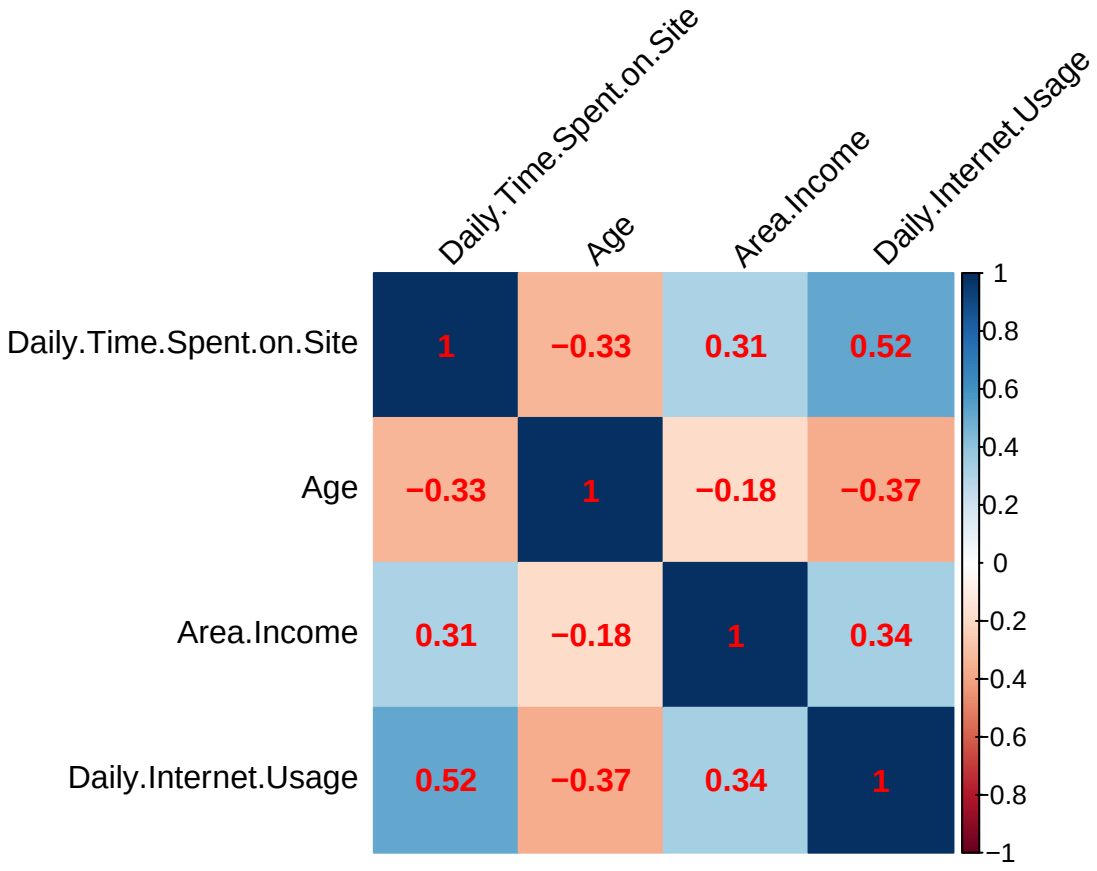
```
# Gender who visited the site most
ggplot(data = df1) +
  geom_bar(mapping = aes(x = Male))
```




```
# 0= F 1= M we see more females visited the site
```

BIvariate Analysis

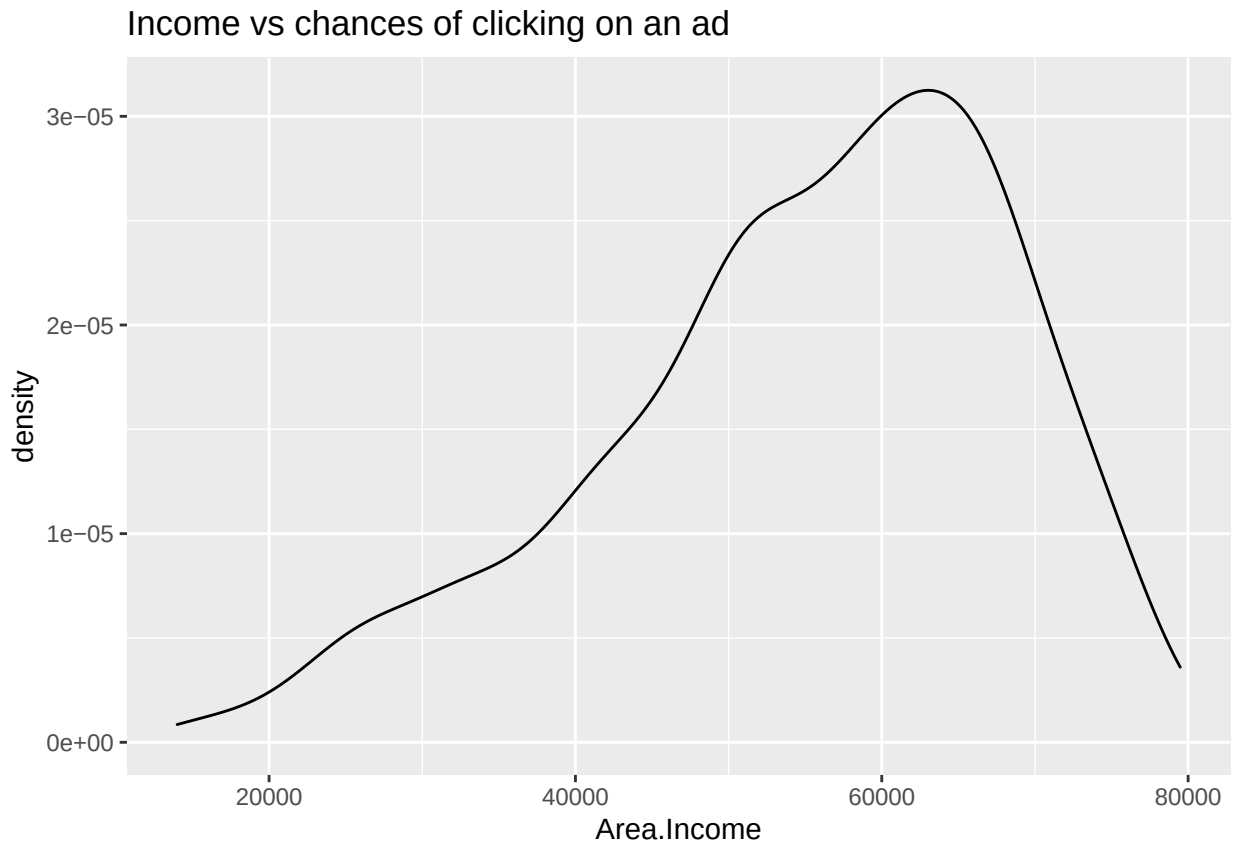
```
# correlation plot
res = cor(num)
corrplot(res, method="color",addCoef.col = "red",
         tl.col="black", tl.srt=45)
```



Observations

There are no strong correlations on either the negative or te positive.

```
# Lets show the relationship between clicking an ad and the people with an income
ggplot(df1,
       aes(x = Area.Income,
           fill = Clicked.on.Ad)) +
  geom_density(alpha = 0.4) +
  labs(title = "Income vs chances of clicking on an ad")
```

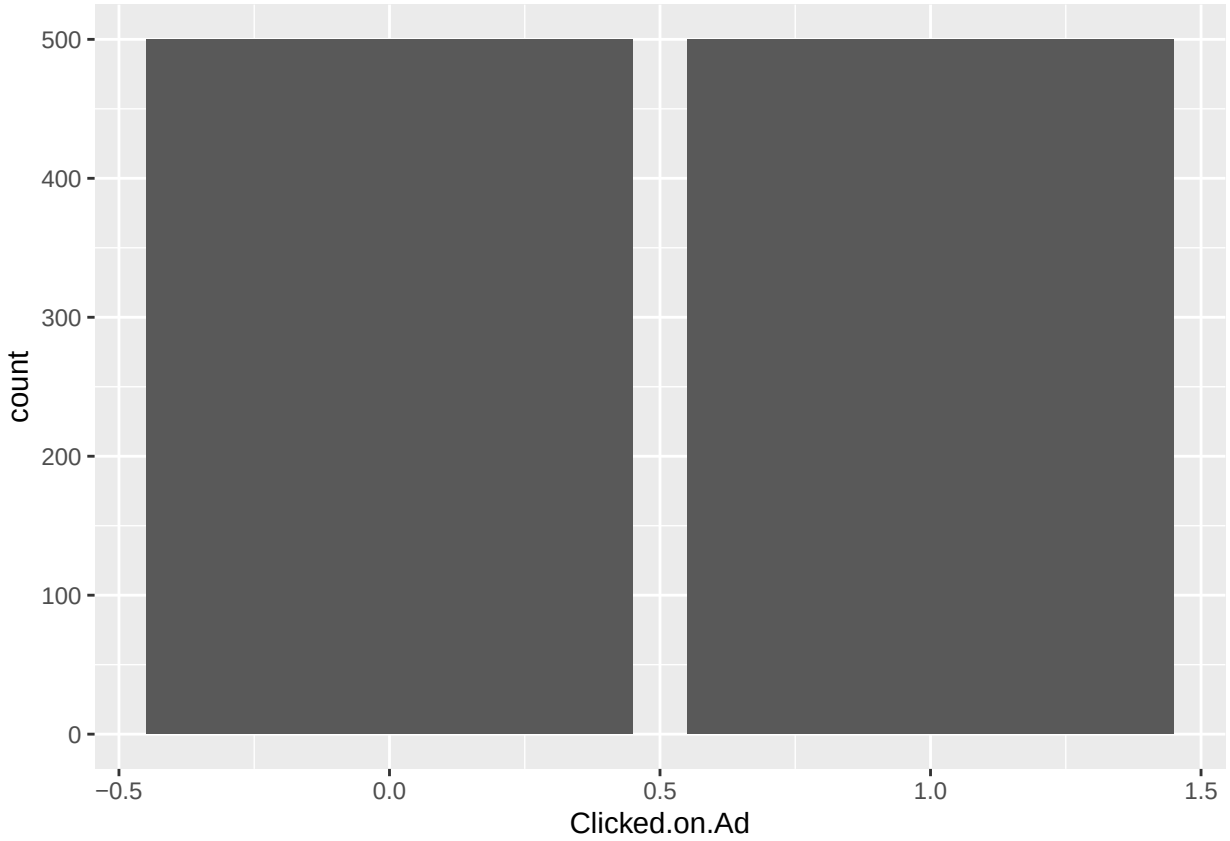


Observations

income from below 60000 are likely to click on an ad.

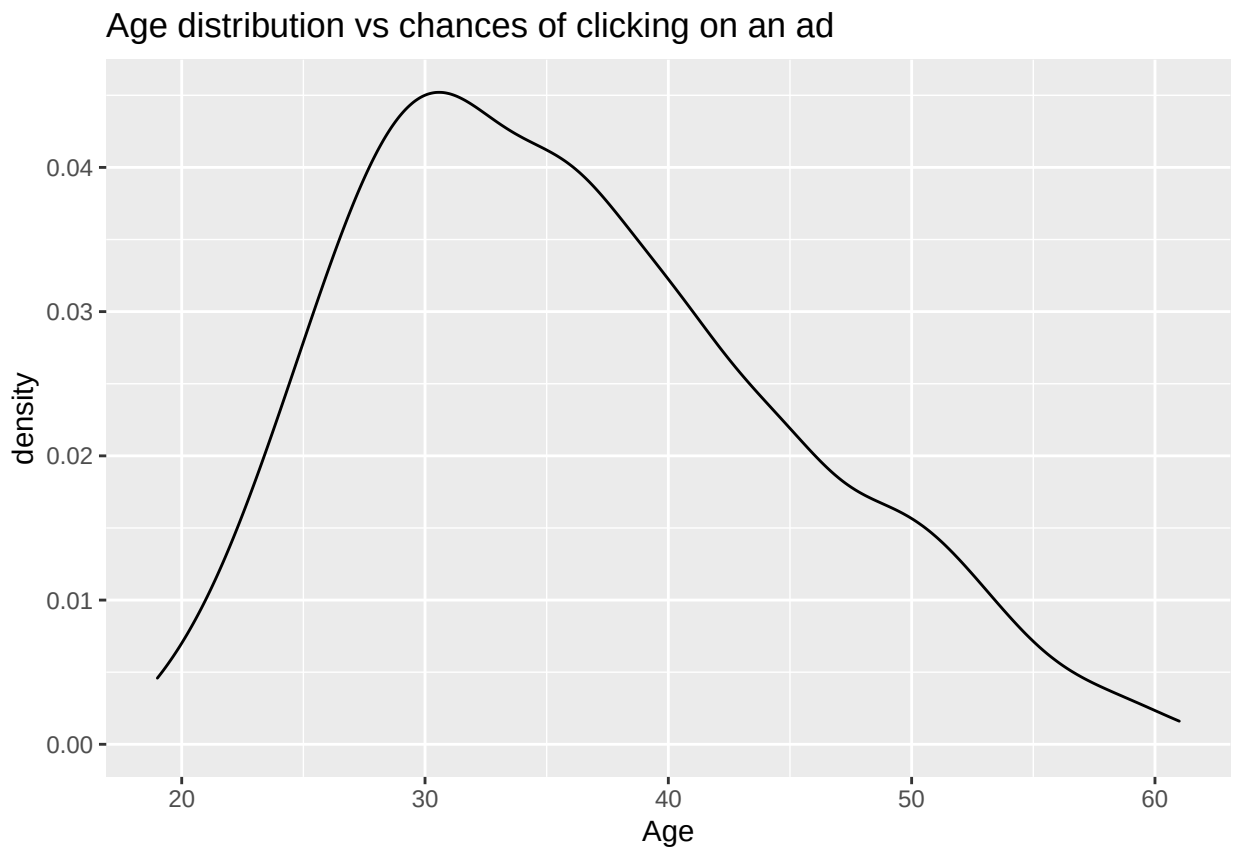
There is a likelihood on all income class clicking the ad however.

```
# Lets compar ethe gender that is likely to click on an ad
ggplot(df1,
       aes(x = Clicked.on.Ad,
           fill = Male)) +
  geom_bar(position = "stack")
```



Observation The males are most frequent to clicking ads unlike women

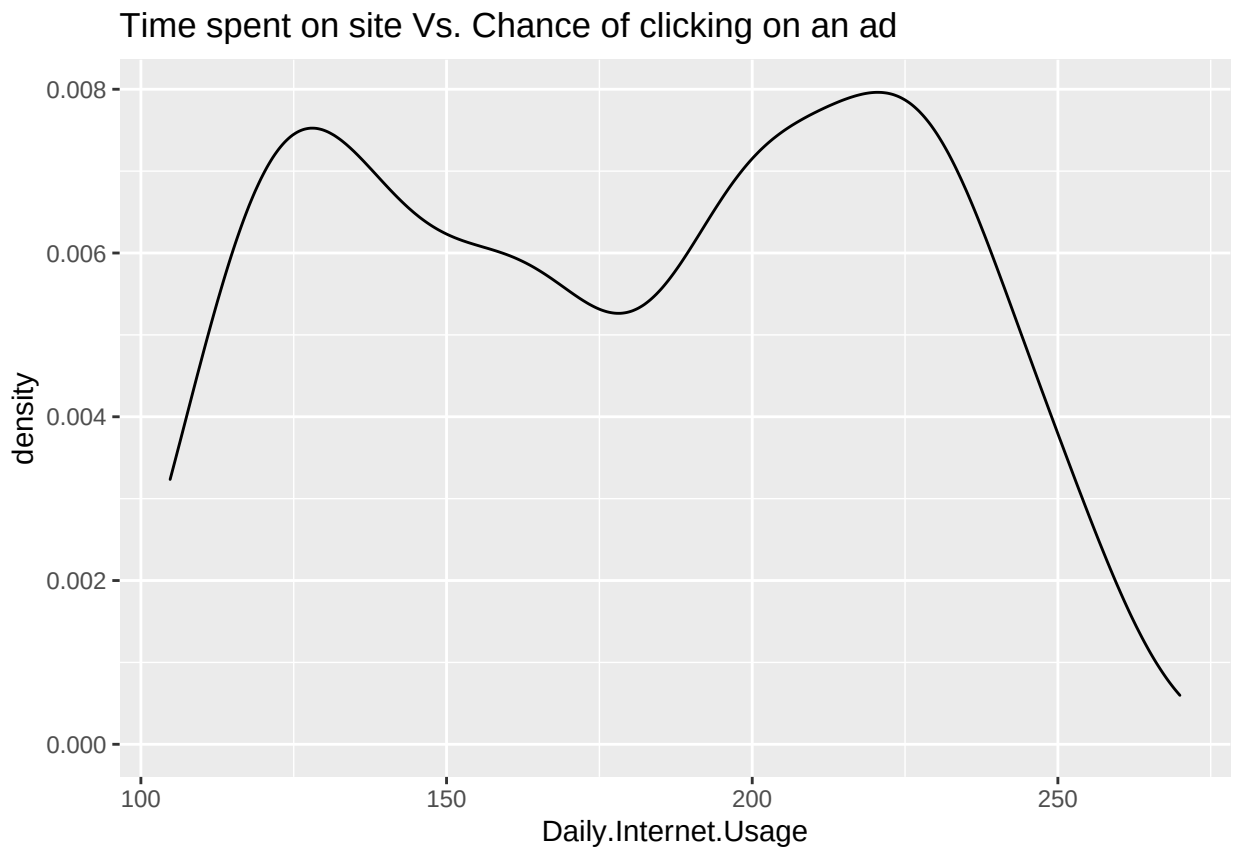
```
# lets observe the age and the likelihood of clicking an ad
ggplot(df1,
       aes(x = Age,
           fill = Clicked.on.Ad)) +
  geom_density(alpha = 0.4) +
  labs(title = "Age distribution vs chances of clicking on an ad")
```



Observation

All ages seem to click on the ads, but the age between 30 and 45 have the highest chance.

```
# Lets see Internet Usage and it's relationship to clicking an ad
ggplot(df1,
       aes(x = Daily.Internet.Usage,
           fill = Clicked.on.Ad)) +
  geom_density(alpha = 0.4) +
  labs(title = "Time spent on site Vs. Chance of clicking on an ad")
```



Observations People on the net less hours click on ads often unlike people who spend most time on the net.

Conclusions

. The males are most frequent to clicking ads unlike women, however the female spend more time on the internet comapared to the male.

. The results show those who spend less time are the most likely to click the ads therefore or male gender would be a suitable target

. The income levels, one with less than 60000, would be a suitable target they are probably looking for a way to add more coins to their pockets

Challenging the solution

This insights were particularly drawn from Eda analysis, therefore need to apply further technics would help give better and conclusive results.