Advertising IP

Jenipher Mawia

10/29/2020

1. Defining the question

Perform extensive data cleaning and Exploratory Data Analysis on the following data and provide relevant conclusion and recommendation. Also build a model using supervised learning algorithms to predict whether a user will click on the ad or not

1.1 Specifying the 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.
- Build a supervised learning model to make the prediction

2. Defining the metrics for success

This project will be considered a success if:

- the above named specific questions are answered/accomplished

3. 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.

4. Experimental Design Taken

The following is the order in which I went about this project:

- Data Sourcing and Understanding
- Checking the data (head and tail, shape(number of records), datatypes)
- Data cleaning procedures (handling null values,outliers, anomalies)

- Exploratory data analysis (Univariate, Bivariate and Multivariate analyses)
- Implementing the solution
- Challenging the solution
- Conclusion and recommendation

5. Data Understanding

Reading the data

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

Checking the data

Shape

```
dim(advertising)
```

```
## [1] 1000   10
```

Head and Tail of the data

Head

```
# checking the first 6 rows in 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
```

Tail

```
# checking the last 6 rows in 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
```

Data Types

```
#checking the datatypes of the columns
str(advertising)
```

```
## '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 standardization" "Organic bottom-line service-desk"
"Triple-buffered reciprocal time-frame" ...
##  $ 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" "2016-01-10 02:31:19" ...
##  $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

6. Appropriateness of the available data to answer the given question

The data above contains 1000 entries and 10 columns(fields). The data contains numeric and character(string) datatypes. These columns include: “Daily time spent on Site”, “age”, “Daily internet usage”, “country”, “gender”, “clicked on ad”(Y/N) etc.

All these fields can be used to determine the patterns of clients/customers and help to identify which individuals are most likely to click on ads.

Therefore, it can be concluded that the data available is appropriate and relevant to answer the given question.

7. Data Cleaning

Changing the column names format

From the above outputs, we can see that the column names are not in the appropriate formats which needs to be changed.

```
# get column names
colnames(advertising)
```

```
##  [1] "Daily.Time.Spent.on.Site" "Age"
##  [3] "Area.Income"              "Daily.Internet.Usage"
##  [5] "Ad.Topic.Line"            "City"
##  [7] "Male"                     "Country"
##  [9] "Timestamp"                "Clicked.on.Ad"
```

```
# rename the column names
names(advertising)[names(advertising) == "Daily.Time.Spent.on.Site"] <-
"daily_time_spent_on_site"
names(advertising)[names(advertising) == "Age"] <- "age"
names(advertising)[names(advertising) == "Area.Income"] <- "area_income"
names(advertising)[names(advertising) == "Daily.Internet.Usage"] <-
"daily_internet_usage"
names(advertising)[names(advertising) == "Ad.Topic.Line"] <- "ad_topic_line"
names(advertising)[names(advertising) == "City"] <- "city"
names(advertising)[names(advertising) == "Male"] <- "male"
names(advertising)[names(advertising) == "Country"] <- "country"
names(advertising)[names(advertising) == "Timestamp"] <- "timestamp"
names(advertising)[names(advertising) == "Clicked.on.Ad"] <- "clicked_on_ad"
```

```
# preview changes made
colnames(advertising)
```

```
##  [1] "daily_time_spent_on_site" "age"
##  [3] "area_income"              "daily_internet_usage"
##  [5] "ad_topic_line"            "city"
```

```
##  [7] "male"                      "country"
##  [9] "timestamp"                 "clicked_on_ad"
```

Missing data

```
#check for missing values in the data per 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 aren't any missing values in the data

Duplicate entries

```
# check for any duplicate entries
duplicates <- advertising[duplicated(advertising),]
duplicates
```

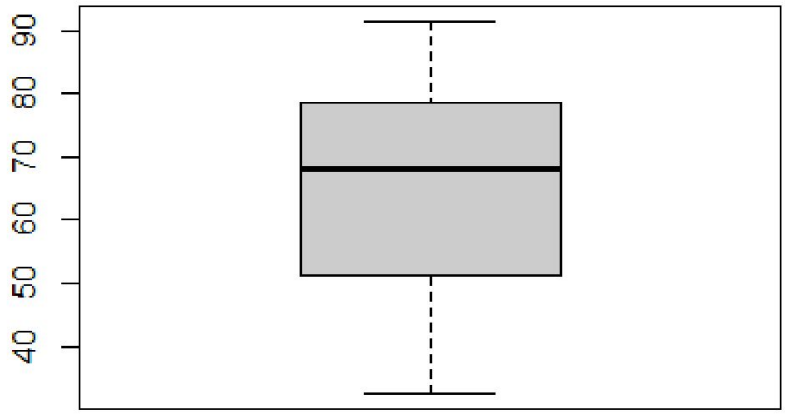
```
##  [1] daily_time_spent_on_site age                      area_income
##  [4] daily_internet_usage     ad_topic_line            city
##  [7] male                     country                  timestamp
## [10] clicked_on_ad
## <0 rows> (or 0-length row.names)
```

There aren't any duplicated entries in the data

Outliers

Check for outliers.

```
boxplot(advertising$daily_time_spent_on_site)
```



Get numerical columns to check for outliers from

```
# check which of the columns has numeric data
nums <- unlist(lapply(advertising, is.numeric))
nums
```

```
## daily_time_spent_on_site                       age               area_income
##                     TRUE                      TRUE                      TRUE
##     daily_internet_usage            ad_topic_line                      city
##                     TRUE                     FALSE                     FALSE
##                     male                   country                 timestamp
##                     TRUE                     FALSE                     FALSE
##            clicked_on_ad
##                     TRUE
```

```
# output the numeric columns in form of a dataframe and check the top of the
resulting dataframe
numerical <- advertising[ , nums]
head(numerical)
```

```
##   daily_time_spent_on_site age area_income daily_internet_usage male
## 1                    68.95  35    61833.90               256.09    0
## 2                    80.23  31    68441.85               193.77    1
## 3                    69.47  26    59785.94               236.50    0
## 4                    74.15  29    54806.18               245.89    1
## 5                    68.37  35    73889.99               225.58    0
## 6                    59.99  23    59761.56               226.74    1
##   clicked_on_ad
```

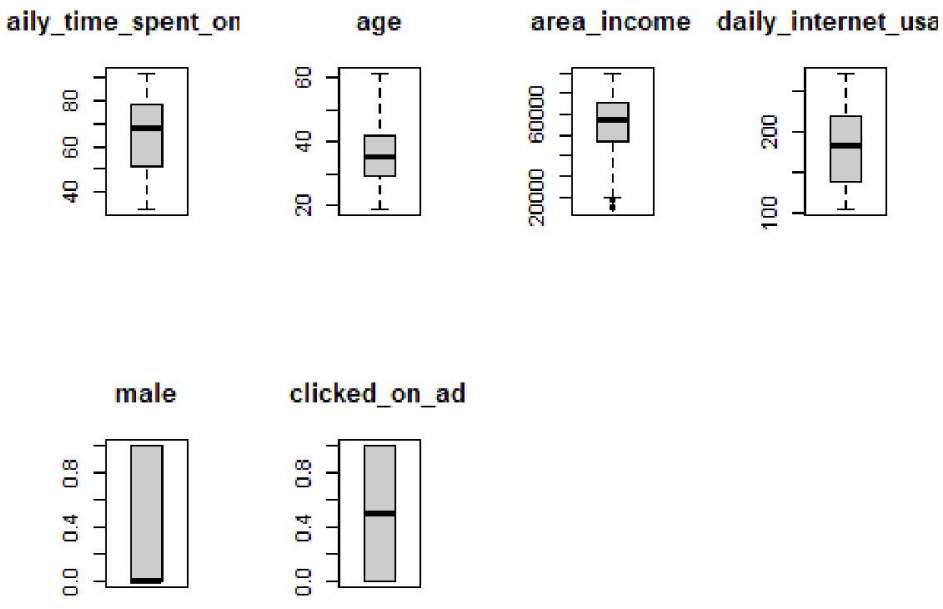
```
## 1              0
## 2              0
## 3              0
## 4              0
## 5              0
## 6              0
```

```
#advertising[ , purrr::map_lgl(advertising, is.numeric)]

#dplyr::select_if(advertising, is.numeric)
```

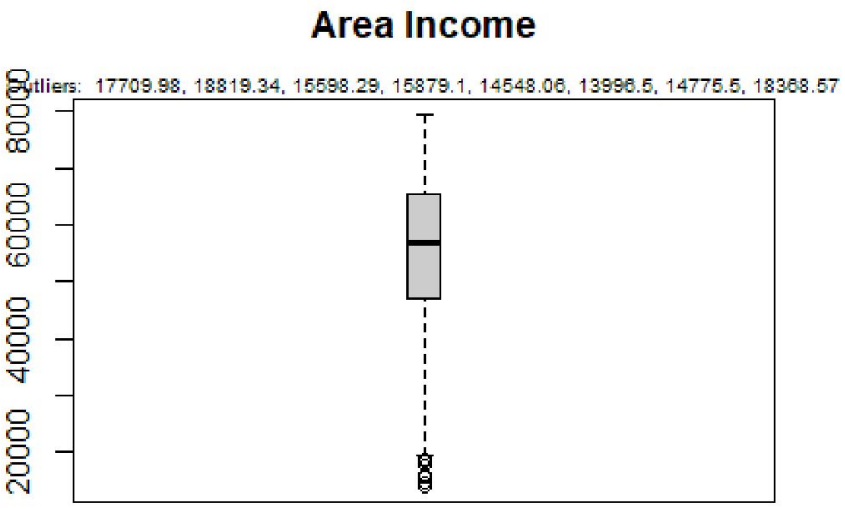
Only 6 columns out of the total 10 are numeric. The rest contain non-numeric data.

```
# make multiple boxplots of the numerical columns to check for any outliers
present
par(mfrow=c(2, 4))
for (i in 1:length(numerical)) {
        boxplot(numerical[,i], main=names(numerical[i]), type="l")
}
```



There are a few outliers present in the column “area_income”.

```
### outlier values in the area_income column
outlier_values <- boxplot.stats(advertising$area_income)$out
boxplot(advertising$area_income, main="Area Income", boxwex=0.1)
mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=0.6)
```



Dealing with outliers

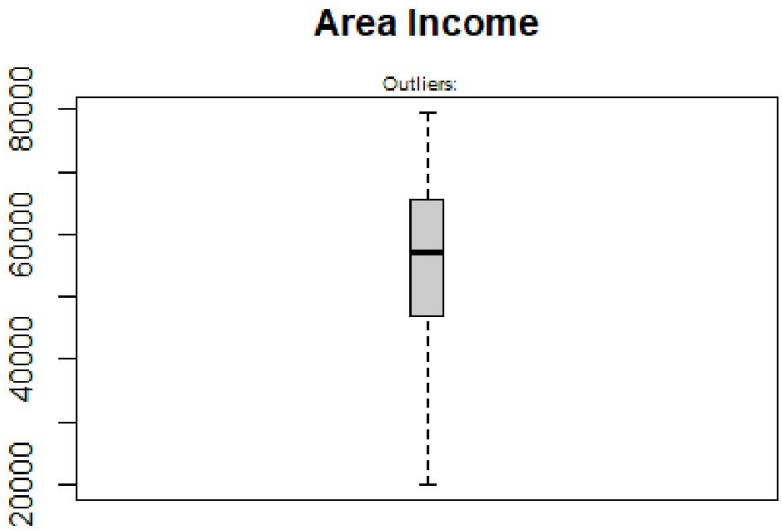
There are various ways of dealing with outliers:

Capping

```
# capping
#x <- advertising$Area.Income
qnt <- quantile(advertising$area_income, probs=c(.25, .75), na.rm = T)
caps <- quantile(advertising$area_income, probs=c(.05, .95), na.rm = T)
H <- 1.5 * IQR(advertising$area_income, na.rm = T)
advertising$area_income[advertising$area_income < (qnt[1] - H)] <- caps[1]
advertising$area_income[advertising$area_income > (qnt[2] + H)] <- caps[2]
```

make a boxplot of the Area.Income column to see the changes made

```
### outlier values in the Area.Income column
outlier_values <- boxplot.stats(advertising$area_income)$out
boxplot(advertising$area_income, main="Area Income", boxwex=0.1)
mtext(paste("Outliers: ", paste(outlier_values, collapse=", ")), cex=0.6)
```



Now we can make a plot of all numerical columns in the data once more to ensure no more outliers are present

```
# reassign the "advertising" dataframe onto a new variable to avoid
corrupting the original data
data <-advertising

#outliers <- boxplot(advertising$Area.Income, plot=FALSE)$out
#data <- data[-which(data$Area.Income %in% outliers),]
```

Getting numerical columns

```
nums1 <- unlist(lapply(data, is.numeric))

# output the numeric columns in form of a dataframe and check the top of the
resulting dataframe
numericals <- data[ , nums]
head(numericals)
```

```
##   daily_time_spent_on_site age area_income daily_internet_usage male
## 1                    68.95  35    61833.90                256.09    0
## 2                    80.23  31    68441.85                193.77    1
## 3                    69.47  26    59785.94                236.50    0
## 4                    74.15  29    54806.18                245.89    1
## 5                    68.37  35    73889.99                225.58    0
## 6                    59.99  23    59761.56                226.74    1
##   clicked_on_ad
## 1             0
```

```
## 2              0
## 3              0
## 4              0
## 5              0
## 6              0
```

Plotting

```
par(mfrow=c(2, 4))
for (i in 1:length(numericals)) {
        boxplot(numericals[,i], main=names(numericals[i]), type="l")
}
```

No more outliers are present in the data.

Anomalies

Anomalies are inconsistencies in the data and this can be checked for in many ways. These are rare items, events or observations which raise suspicions by differing significantly from the majority of the data.

Data-Type Conversion

```
# checking the datatypes of each column
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 standardization" "Organic bottom-line service-desk"
"Triple-buffered reciprocal time-frame" ...
##  $ 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" "2016-01-10 02:31:19" ...
##  $ clicked_on_ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

The columns “male”(representing the gender of the client-given by values 0 and 1) and “clicked_on_ad”(Y/N values represented by 0 and 1) are categorical values. It is possible to convert them to factor type so that they can have only two levels.

The “timestamp” column also requires to be converted into date-time format

```
# change the datatypes of the two columns
data$male <- as.factor(data$male)
data$clicked_on_ad <- as.factor(data$clicked_on_ad)

# check if the "male" column is a factor
is.factor(data$male)
```

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

```
# create a temporary dataframe containing the data
temp <- data
library(anytime)
# converting the datatype of the column "timestamp"
temp$timestamp <- anytime::anydate(temp$timestamp)
# check the datatype of the column
str(temp$timestamp)
```

```
##  Date[1:1000], format: "2016-03-27" "2016-04-04" "2016-03-13" "2016-01-10"
"2016-06-03" ...
```

As we can see above the anydate() function converts the characters that it recognizes to be part of a date into a date class and ignores all other characters in the string(the time function). We use the POSIXCt function instead

```
# converting the datatype of the column "timestamp" on the original data
data$timestamp <- as.POSIXct(data$timestamp, format="%Y-%m-%d %H:%M:%S")
str(data$timestamp)
```

```
##  POSIXct[1:1000], format: "2016-03-27 00:53:11" "2016-04-04 01:39:02"
"2016-03-13 20:35:42" ...
```

Then extract the year, month, day and hour from the timestamp column. The minute and second functions of time are not as important in the analysis.

```
# extract the year, month, day and hour from the timestamp column
data$year <- format(data$timestamp, format="%Y")
data$month <- format(data$timestamp, format="%m")
data$day <- format(data$timestamp, format="%d")
data$hour <- format(data$timestamp, format="%H")

str(data)
```

```
## 'data.frame':    1000 obs. of  14 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 standardization" "Organic bottom-line service-desk"
"Triple-buffered reciprocal time-frame" ...
##  $ city                    : chr  "Wrightburgh" "West Jodi" "Davidton"
"West Terrifurt" ...
##  $ male                    : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 2
2 ...
##  $ country                 : chr  "Tunisia" "Nauru" "San Marino" "Italy"
...
##  $ timestamp               : POSIXct, format: "2016-03-27 00:53:11"
"2016-04-04 01:39:02" ...
##  $ clicked_on_ad           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1
1 ...
##  $ year                    : chr  "2016" "2016" "2016" "2016" ...
##  $ month                   : chr  "03" "04" "03" "01" ...
##  $ day                     : chr  "27" "04" "13" "10" ...
##  $ hour                    : chr  "00" "01" "20" "02" ...
```

```
#convert the new columns created to categorical values(factor)
data$year <- as.factor(data$year)
data$month <- as.factor(data$month)
data$day <- as.factor(data$day)
data$hour <- as.factor(data$hour)
#check the datatypes of the resulting dataframe
str(data)
```

```
## 'data.frame':    1000 obs. of  14 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 standardization" "Organic bottom-line service-desk"
"Triple-buffered reciprocal time-frame" ...
##  $ city                    : chr  "Wrightburgh" "West Jodi" "Davidton"
"West Terrifurt" ...
##  $ male                    : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 2
2 ...
##  $ country                 : chr  "Tunisia" "Nauru" "San Marino" "Italy"
...
##  $ timestamp               : POSIXct, format: "2016-03-27 00:53:11"
"2016-04-04 01:39:02" ...
##  $ clicked_on_ad           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1
1 ...
##  $ year                    : Factor w/ 1 level "2016": 1 1 1 1 1 1 1 1 1 1
...
##  $ month                   : Factor w/ 7 levels "01","02","03",..: 3 4 3 1
6 5 1 3 4 7 ...
##  $ day                     : Factor w/ 31 levels "01","02","03",..: 27 4
13 10 3 19 28 7 18 11 ...
##  $ hour                    : Factor w/ 24 levels "00","01","02",..: 1 2 21
3 4 15 21 2 10 2 ...
```

- The “year” column has only one level;2016. This means the data was collected in the year 2016.
- The “month” column has 7 levels; months January to July.
- The “day” column is a factor of 31 levels indicating that the number of days represented are 31.
- The “hour” column is also a factor of 24 levels indicating the number of hours in a day.

We can now delete the timestamp column as we do not need it anymore and move the column “clicked_on_add” to the end(make it the last column in the data)

```
# drop the timestamp column
data$timestamp <- NULL
colnames(data)
```

```
##  [1] "daily_time_spent_on_site" "age"
##  [3] "area_income"              "daily_internet_usage"
##  [5] "ad_topic_line"            "city"
##  [7] "male"                     "country"
##  [9] "clicked_on_ad"            "year"
## [11] "month"                    "day"
## [13] "hour"
```

```
# move the 'clicked_on_ad' column to the end
data <- data[, c(1:8, 10:13, 9)]
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
year
## 1    Cloned 5thgeneration orchestration    Wrightburgh    0    Tunisia
2016
## 2    Monitored national standardization      West Jodi    1      Nauru
2016
## 3      Organic bottom-line service-desk       Davidton    0 San Marino
2016
## 4 Triple-buffered reciprocal time-frame West Terrifurt    1      Italy
2016
## 5         Robust logistical utilization   South Manuel    0    Iceland
2016
## 6       Sharable client-driven software      Jamieberg    1     Norway
2016
##   month day hour clicked_on_ad
## 1    03  27   00             0
## 2    04  04   01             0
## 3    03  13   20             0
## 4    01  10   02             0
## 5    06  03   03             0
## 6    05  19   14             0
```

8. Exploratory Data Analysis

8.1 Univariate Data Analysis

Measures of Central Tendency

1. Mean

Get the mean of each numerical column

```
# here we use the dataframe "numericals" initially created when plotting
boxplots, containing only numeric columns
colMeans(numericals)
```

```
## daily_time_spent_on_site                      age              area_income
##                  65.0002                  36.0090               55105.4371
##     daily_internet_usage                     male            clicked_on_ad
##                180.0001                   0.4810                   0.5000
```

2. Median

Get the median of each numerical column

```
apply(numericals,2,median)
```

```
## daily_time_spent_on_site                      age              area_income
##                   68.215                   35.000                57012.300
##     daily_internet_usage                     male            clicked_on_ad
##                  183.130                    0.000                    0.500
```

3. Mode

Get the mode of each numerical column

Daily time spent on site

```
# Create the function.
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Calculate the mode using the user function.
daily_time_on_site_mode <- getmode(data$daily_time_spent_on_site)
print(daily_time_on_site_mode)
```

```
## [1] 62.26
```

Most users spent at least 62.26 minutes on the site.

Age

```
age_mode <- getmode(data$age)
print(age_mode)
```

```
## [1] 31
```

A large number of users visiting the site are of 31 years of age

Area Income

```
area_income_mode <- getmode(data$area_income)
print(area_income_mode)
```

```
## [1] 28275.3
```

Most users visiting the site have an area income of 28275.3

Daily Internet Usage

```
daily_internet_usage_mode <- getmode(data$daily_internet_usage)
print(daily_internet_usage_mode)
```

```
## [1] 167.22
```

Daily internet usage for most users visiting the site is 167.22

Ad Topic Line

```
ad_topic_line_mode <- getmode(data$ad_topic_line)
print(ad_topic_line_mode)
```

```
## [1] "Cloned 5thgeneration orchestration"
```

The most frequent Ad Topic line is “Cloned 5thgeneration orchestration”

City

```
city_mode <- getmode(data$city)
print(city_mode)
```

```
## [1] "Lisamouth"
```

The most popular city is “Lisamouth”

Gender

```
male_mode <- getmode(data$male)
print(male_mode)
```

```
## [1] 0
## Levels: 0 1
```

Most users visiting the site are female

Country

```
country_mode <- getmode(data$country)
print(country_mode)
```

```
## [1] "Czech Republic"
```

Most users visiting the site are from the country Czech Republic

Year

```
year_mode <- getmode(data$year)
print(year_mode)
```

```
## [1] 2016
## Levels: 2016
```

The year column is a factor of 1 level: year 2016. The data was collected in 2016.

Month

```
month_mode <- getmode(data$month)
print(month_mode)
```

```
## [1] 02
## Levels: 01 02 03 04 05 06 07
```

The modal month is February.

Day

```
day_mode <- getmode(data$day)
print(day_mode)
```

```
## [1] 03
## 31 Levels: 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 ...
31
```

Most users visited the site on the third day of the month.

Hour

```
hour_mode <- getmode(data$hour)
print(hour_mode)
```

```
## [1] 07
## 24 Levels: 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 ...
23
```

The most popular hour that users visit the site is 0700hrs.

Clicked on ad

```
clicked_on_ad_mode <- getmode(data$clicked_on_ad)
print(clicked_on_ad_mode)
```

```
## [1] 0
## Levels: 0 1
```

Most users visiting the site did not click on the ad

Measures of Dispersion

- Find the **minimum, maximum and quantiles** of the columns in the data.

```
summary(data)
```

```
##  daily_time_spent_on_site      age         area_income
daily_internet_usage
##  Min.   :32.60            Min.   :19.00   Min.   :19992   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   :55105   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             year
##  Length:1000        Length:1000        0:519   Length:1000
```

```
2016:1000
##  Class :character   Class :character   1:481   Class :character
##  Mode  :character   Mode  :character           Mode  :character
##
##
##
##
##  month         day           hour      clicked_on_ad
##  01:147   03     : 46   07     : 54   0:500
##  02:160   17     : 42   20     : 50   1:500
##  03:156   15     : 41   09     : 49
##  04:147   10     : 37   21     : 48
##  05:147   04     : 36   00     : 45
##  06:142   26     : 36   05     : 44
##  07:101   (Other):762   (Other):710
```

- **Range**

Daily Time Spent on the site

```
range(data$daily_time_spent_on_site)
```

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

The time spent by most users visiting the site is between 32.6-91.43 minutes

Age

```
range(data$age)
```

```
## [1] 19 61
```

Users visiting the site are adults between ages 19-61.

Area Income

```
range(data$area_income)
```

```
## [1] 19991.72 79484.80
```

Area incomes for users visiting the site is between 19000 and 79484

Daily Internet Usage

```
range(data$daily_internet_usage)
```

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

Users visiting the site use data bundles of ranges between 104.78-269.96 on a daily basis.

- **Interquartile Range**

The interquartile range also commonly known as IQR is the range between the 1st and 3rd quantiles. It is the difference between the two quantiles.

Daily time spent on site

```
IQR(data$daily_time_spent_on_site)
```

```
## [1] 27.1875
```

Age

```
IQR(data$age)
```

```
## [1] 13
```

Area Income

```
IQR(data$area_income)
```

```
## [1] 18438.83
```

Daily Internet Usage

```
IQR(data$daily_internet_usage)
```

```
## [1] 79.9625
```

- **Standard Deviation**

Find the standard deviation of the various columns in the data

```
 apply(numericals,2,sd)
```

```
## daily_time_spent_on_site                      age              area_income
##             1.585361e+01             8.785562e+00             1.315412e+04
##     daily_internet_usage                     male            clicked_on_ad
##             4.390234e+01             4.998889e-01             5.002502e-01
```

- **Variance**

Find the variance of the numerical columns

```
sapply(numericals, var)
```

```
## daily_time_spent_on_site                      age              area_income
##             2.513371e+02             7.718611e+01             1.730310e+08
##     daily_internet_usage                     male            clicked_on_ad
##             1.927415e+03             2.498889e-01             2.502503e-01
```

- **Kurtosis**

Find the kurtosis of continuos numerical columns in the data

Daily time spent on site

```
library(e1071)
kurtosis(numericals$daily_time_spent_on_site)
```

```
## [1] -1.099864
```

The kurtosis for this variable is less than 3 implying that the distribution of this variable is platykurtic. This means that there are few to no outliers which we have observed above when dealing with outliers.

Age

```
kurtosis(numericals$age)
```

```
## [1] -0.4097066
```

The distribution is platykurtic implying the existence of few to no outliers.

Area Income

```
kurtosis(numericals$area_income)
```

```
## [1] -0.3703758
```

A kurtosis value of 2.63 indicates that the distribution is platykurtic although very close to being mesokurtic. It exhibits presence of outliers as observed above from the boxplots.

Daily Internet Usage

```
kurtosis(numericals$daily_internet_usage)
```

```
## [1] -1.275752
```

The distribution is platykurtic.

Gender

```
kurtosis(numericals$male)
```

```
## [1] -1.996226
```

Clicked on ad

```
kurtosis(numericals$clicked_on_ad)
```

```
## [1] -2.001999
```

- **Skewness**

Find the skewness of all continuous numerical columns

Daily time spent on site

```
library(e1071)
skewness(data$daily_time_spent_on_site)
```

```
## [1] -0.370646
```

This proves that this variable is slightly negatively skewed(the distribution is skewed to the left).

Age

```
skewness(data$age)
```

```
## [1] 0.4777052
```

This skewness value implies that the distribution is almost fairly symmetrical

Area Income

```
skewness(data$area_income)
```

```
## [1] -0.560965
```

The distribution is negatively skewed.

Daily Internet Usage

```
skewness(data$daily_internet_usage)
```

```
## [1] -0.03343681
```

The distribution is negatively skewed but by a very small value close to 0.

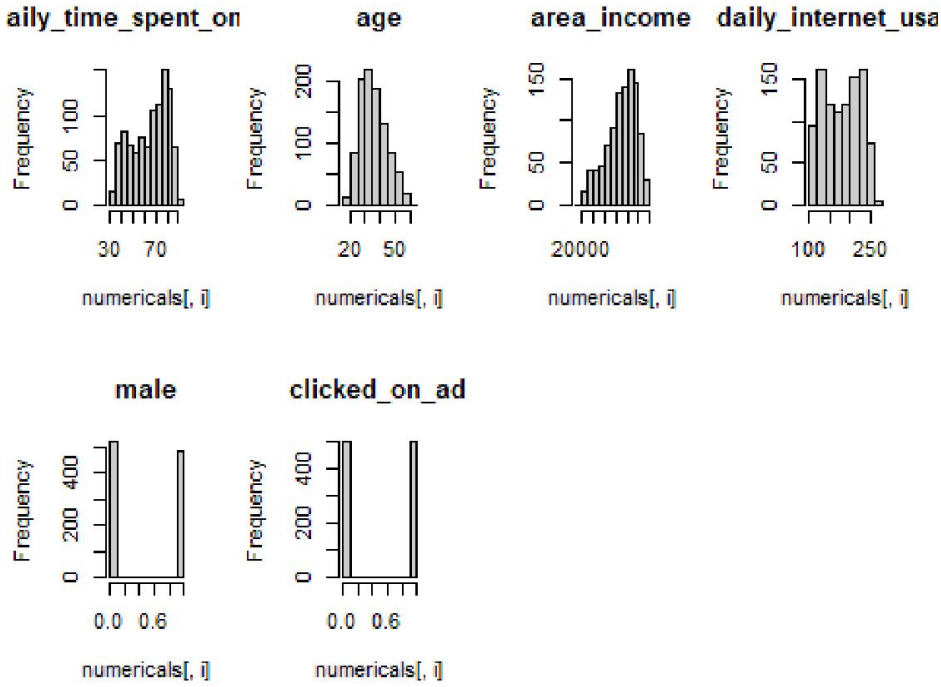
The skewness of the various numerical columns can be observed by checking the distribution of the data using histograms.

```
#colkurtosis(numericals)
```

```
#colskewness(numericals, pvalue = FALSE)
```

Histograms

```
par(mfrow=c(2, 4))
for (i in 1:length(numericals)) {
        hist(numericals[,i], main=names(numericals[i]))
}
```



8.2 Bivariate and Multivariate Analysis

We will investigate the relationship between the target variable("clicked on ad") and the other columns

```
# how many males and females clicked on ads
gender_ad <- table(data$clicked_on_ad, data$male)
names(dimnames(gender_ad)) <- c("Clicked on Ad?", "Male")
gender_ad
```

```
##                Male
## Clicked on Ad?   0   1
##              0 250 250
##              1 269 231
```

The data is not unbalanced. The number of males and females who did not click on an ad are equal. However, more females clicked on the ads compared to males but only by a smaller number

```
# ad clicked per month
month_ad <- table(data$month, data$clicked_on_ad)
names(dimnames(month_ad)) <- c("Month", "Clicked on Ad?")
month_ad
```

```
##       Clicked on Ad?
## Month  0  1
##    01 78 69
##    02 77 83
```

```
##    03 82 74
##    04 73 74
##    05 68 79
##    06 71 71
##    07 51 50
```

We can see that February reports the highest number of ads clicked and July the least.

```
# ad clicked per day
day_ad <- table(data$day, data$clicked_on_ad)
names(dimnames(day_ad)) <- c("Day", "Clicked on Ad?")
day_ad
```

```
##     Clicked on Ad?
## Day   0  1
##   01 14 19
##   02 15 10
##   03 20 26
##   04 22 14
##   05 17 18
##   06 11 14
##   07 18 14
##   08 20 15
##   09 14 20
##   10 18 19
##   11 17 15
##   12  9 20
##   13 13 17
##   14 12 21
##   15 21 20
##   16 21 14
##   17 24 18
##   18 18 17
##   19 17 12
##   20 22 11
##   21 17 15
##   22 14 10
##   23 13 22
##   24 15 18
##   25  8 15
##   26 21 15
##   27 19 16
##   28 13 17
##   29 14 15
##   30 14 14
##   31  9  9
```

The 3rd day of the month reports the highest record of users clicking ads while the 31st day reports the lowest number of visitors to the site.

```r
# ad clicked per hour
hour_ad <- table(data$hour, data$clicked_on_ad)
names(dimnames(hour_ad)) <- c("Hour", "Clicked on Ad?")
hour_ad
```

```
##     Clicked on Ad?
## Hour  0  1
##   00 19 26
##   01 16 16
##   02 19 17
##   03 19 23
##   04 21 21
##   05 23 21
##   06 16 23
##   07 28 26
##   08 22 21
##   09 21 28
##   10 17 14
##   11 16 24
##   12 22 16
##   13 21 21
##   14 22 21
##   15 16 19
##   16 23 16
##   17 18 23
##   18 16 25
##   19 20 19
##   20 26 24
##   21 29 19
##   22 24 19
##   23 26 18
```

At 9am, most users clicked on the ad while at 10am very few users clicked on the ads. It could be that at 10am users are so engrossed in their daily work.

```r
# ad clicked per country
country_ad <- table(data$country, data$clicked_on_ad)
names(dimnames(country_ad)) <- c("Country", "Clicked on Ad")
country_ad
```

```
##                                                  Clicked on Ad
## Country                                           0 1
##   Afghanistan                                     3 5
##   Albania                                         3 4
##   Algeria                                         3 3
##   American Samoa                                  2 3
##   Andorra                                         0 2
##   Angola                                          3 1
##   Anguilla                                        3 3
##   Antarctica (the territory South of 60 deg S)    1 2
##   Antigua and Barbuda                             1 4
```

```
##   Argentina                                                1 1
##   Armenia                                                  2 1
##   Aruba                                                    1 0
##   Australia                                                1 7
##   Austria                                                  4 1
##   Azerbaijan                                               2 1
##   Bahamas                                                  3 4
##   Bahrain                                                  3 2
##   Bangladesh                                               2 2
##   Barbados                                                 3 2
##   Belarus                                                  3 3
##   Belgium                                                  3 2
##   Belize                                                   2 3
##   Benin                                                    1 1
##   Bermuda                                                  1 0
##   Bhutan                                                   1 1
##   Bolivia                                                  6 0
##   Bosnia and Herzegovina                                   4 3
##   Bouvet Island (Bouvetoya)                                3 2
##   Brazil                                                   2 3
##   British Indian Ocean Territory (Chagos Archipelago) 0 1
##   British Virgin Islands                                   2 1
##   Brunei Darussalam                                        3 2
##   Bulgaria                                                 2 4
##   Burkina Faso                                             3 1
##   Burundi                                                  5 2
##   Cambodia                                                 5 2
##   Cameroon                                                 5 0
##   Canada                                                   2 3
##   Cape Verde                                               1 0
##   Cayman Islands                                           2 3
##   Central African Republic                                 1 1
##   Chad                                                     2 2
##   Chile                                                    1 3
##   China                                                    2 4
##   Christmas Island                                         2 4
##   Colombia                                                 1 1
##   Comoros                                                  1 1
##   Congo                                                    1 3
##   Cook Islands                                             2 1
##   Costa Rica                                               4 2
##   Cote d'Ivoire                                            1 3
##   Croatia                                                  6 0
##   Cuba                                                     1 4
##   Cyprus                                                   4 4
##   Czech Republic                                           5 4
##   Denmark                                                  1 2
##   Djibouti                                                 1 1
##   Dominica                                                 3 2
##   Dominican Republic                                       2 2
```

```
##   Ecuador                                          3 2
##   Egypt                                            2 3
##   El Salvador                                      2 4
##   Equatorial Guinea                                1 3
##   Eritrea                                          4 3
##   Estonia                                          2 1
##   Ethiopia                                         0 7
##   Falkland Islands (Malvinas)                      2 2
##   Faroe Islands                                    1 2
##   Fiji                                             4 3
##   Finland                                          4 1
##   France                                           4 5
##   French Guiana                                    1 3
##   French Polynesia                                 4 1
##   French Southern Territories                      4 1
##   Gabon                                            6 0
##   Gambia                                           1 1
##   Georgia                                          2 2
##   Germany                                          0 1
##   Ghana                                            2 2
##   Gibraltar                                        3 0
##   Greece                                           5 3
##   Greenland                                        4 1
##   Grenada                                          2 2
##   Guadeloupe                                       1 1
##   Guam                                             2 2
##   Guatemala                                        1 3
##   Guernsey                                         1 2
##   Guinea                                           1 2
##   Guinea-Bissau                                    1 1
##   Guyana                                           2 3
##   Haiti                                            1 1
##   Heard Island and McDonald Islands                1 2
##   Holy See (Vatican City State)                    2 1
##   Honduras                                         3 2
##   Hong Kong                                        2 4
##   Hungary                                          1 5
##   Iceland                                          2 1
##   India                                            2 0
##   Indonesia                                        2 4
##   Iran                                             2 3
##   Ireland                                          2 1
##   Isle of Man                                      2 1
##   Israel                                           2 2
##   Italy                                            4 1
##   Jamaica                                          3 2
##   Japan                                            2 2
##   Jersey                                           2 4
##   Jordan                                           1 0
##   Kazakhstan                                       2 2
```

```
##   Kenya                                          0 4
##   Kiribati                                       0 1
##   Korea                                          2 3
##   Kuwait                                         1 1
##   Kyrgyz Republic                                5 1
##   Lao People's Democratic Republic               2 2
##   Latvia                                         0 4
##   Lebanon                                        2 4
##   Lesotho                                        1 0
##   Liberia                                        2 6
##   Libyan Arab Jamahiriya                         2 2
##   Liechtenstein                                  0 6
##   Lithuania                                      0 3
##   Luxembourg                                     4 3
##   Macao                                          0 3
##   Macedonia                                      1 1
##   Madagascar                                     4 2
##   Malawi                                         2 2
##   Malaysia                                       3 0
##   Maldives                                       2 2
##   Mali                                           3 1
##   Malta                                          3 3
##   Marshall Islands                               0 1
##   Martinique                                     1 3
##   Mauritania                                     1 1
##   Mauritius                                      3 1
##   Mayotte                                        1 5
##   Mexico                                         2 4
##   Micronesia                                     4 4
##   Moldova                                        4 2
##   Monaco                                         2 1
##   Mongolia                                       2 4
##   Montenegro                                     0 2
##   Montserrat                                     0 1
##   Morocco                                        2 1
##   Mozambique                                     1 0
##   Myanmar                                        4 1
##   Namibia                                        1 1
##   Nauru                                          2 1
##   Nepal                                          3 0
##   Netherlands                                    1 3
##   Netherlands Antilles                           4 2
##   New Caledonia                                  0 2
##   New Zealand                                    2 2
##   Nicaragua                                      3 0
##   Niger                                          1 2
##   Niue                                           3 0
##   Norfolk Island                                 3 2
##   Northern Mariana Islands                       1 2
##   Norway                                         1 1
```

```
##   Pakistan                                          4 1
##   Palau                                             2 2
##   Palestinian Territory                             1 2
##   Panama                                            2 0
##   Papua New Guinea                                  2 3
##   Paraguay                                          2 1
##   Peru                                              3 5
##   Philippines                                       3 3
##   Pitcairn Islands                                  1 1
##   Poland                                            3 3
##   Portugal                                          2 1
##   Puerto Rico                                       3 3
##   Qatar                                             4 2
##   Reunion                                           2 0
##   Romania                                           0 1
##   Russian Federation                                2 1
##   Rwanda                                            3 2
##   Saint Barthelemy                                  0 2
##   Saint Helena                                      3 2
##   Saint Kitts and Nevis                             0 1
##   Saint Lucia                                       1 1
##   Saint Martin                                      2 2
##   Saint Pierre and Miquelon                         2 3
##   Saint Vincent and the Grenadines                  3 3
##   Samoa                                             2 4
##   San Marino                                        2 1
##   Sao Tome and Principe                             0 2
##   Saudi Arabia                                      1 3
##   Senegal                                           3 5
##   Serbia                                            2 3
##   Seychelles                                        2 1
##   Sierra Leone                                      0 2
##   Singapore                                         5 1
##   Slovakia (Slovak Republic)                        2 0
##   Slovenia                                          0 1
##   Somalia                                           3 2
##   South Africa                                      2 6
##   South Georgia and the South Sandwich Islands      1 1
##   Spain                                             0 3
##   Sri Lanka                                         4 0
##   Sudan                                             2 0
##   Suriname                                          1 1
##   Svalbard & Jan Mayen Islands                      2 4
##   Swaziland                                         2 0
##   Sweden                                            3 1
##   Switzerland                                       1 3
##   Syrian Arab Republic                              2 1
##   Taiwan                                            3 4
##   Tajikistan                                        1 2
##   Tanzania                                          2 1
```

```
##   Thailand                                          2 2
##   Timor-Leste                                       4 1
##   Togo                                              2 1
##   Tokelau                                           1 3
##   Tonga                                             3 2
##   Trinidad and Tobago                               1 2
##   Tunisia                                           3 1
##   Turkey                                            1 7
##   Turkmenistan                                      4 2
##   Turks and Caicos Islands                          2 3
##   Tuvalu                                            1 3
##   Uganda                                            0 4
##   Ukraine                                           4 1
##   United Arab Emirates                              3 3
##   United Kingdom                                    1 2
##   United States Minor Outlying Islands              2 2
##   United States of America                          2 3
##   United States Virgin Islands                      2 2
##   Uruguay                                           4 1
##   Uzbekistan                                        1 1
##   Vanuatu                                           5 1
##   Venezuela                                         4 3
##   Vietnam                                           1 2
##   Wallis and Futuna                                 3 1
##   Western Sahara                                    3 4
##   Yemen                                             1 2
##   Zambia                                            1 3
##   Zimbabwe                                          2 4
```

The highest number of users that clicked on the ads from a country is 7 from the countries: Turkey, Ethiopia, Australia. For Ethiopia, all users that visited the site clicked on the ads.

```
# ad clicked per city
city_ads <- table(data$city, data$clicked_on_ad)
names(dimnames(city_ads)) <- c("City", "Clicked on Ad")
city_ads
```

```
##                          Clicked on Ad
## City                      0 1
##   Adamsbury               0 1
##   Adamside                0 1
##   Adamsstad               1 0
##   Alanview                1 0
##   Alexanderfurt           0 1
##   Alexanderview           0 1
##   Alexandrafort           1 0
##   Alexisland              1 0
##   Aliciatown              0 1
##   Alvaradoport            0 1
##   Alvarezland             0 1
```

```
##   Amandafort              0 1
##   Amandahaven             0 1
##   Amandaland              1 0
##   Amyfurt                 1 0
##   Amyhaven                1 0
##   Andersonchester         0 1
##   Andersonfurt            0 1
##   Andersonton             1 0
##   Andrewborough           0 1
##   Andrewmouth             1 0
##   Angelhaven              1 0
##   Anthonyfurt             1 0
##   Ashleychester           1 0
##   Ashleymouth             1 0
##   Austinborough           1 0
##   Austinland              1 0
##   Bakerhaven              1 0
##   Barbershire             1 0
##   Beckton                 1 0
##   Benjaminchester         2 0
##   Bernardton              0 1
##   Bethburgh               0 1
##   Birdshire               1 0
##   Blairborough            0 1
##   Blairville              1 0
##   Blevinstown             0 1
##   Bowenview               1 0
##   Boyerberg               0 1
##   Bradleyborough          1 0
##   Bradleyburgh            0 1
##   Bradleyside             0 1
##   Bradshawborough         1 0
##   Bradyfurt               0 1
##   Brandiland              0 1
##   Brandonbury             0 1
##   Brandonstad             1 0
##   Brandymouth             0 1
##   Brendaburgh             1 0
##   Brendachester           0 1
##   Brianabury              1 0
##   Brianfurt               0 1
##   Brianland               0 1
##   Brittanyborough         0 1
##   Brownbury               1 0
##   Brownport               0 1
##   Brownton                0 1
##   Browntown               0 1
##   Brownview               1 0
##   Bruceburgh              1 0
##   Burgessside             0 1
```

```
##   Butlerfort              0 1
##   Calebberg               1 0
##   Cameronberg             0 1
##   Campbellstad            1 0
##   Cannonbury              1 0
##   Carsonshire             1 0
##   Carterburgh             1 0
##   Carterland              0 1
##   Carterport              1 0
##   Carterton               1 0
##   Cassandratown           1 0
##   Catherinefort           0 1
##   Cervantesshire          0 1
##   Chapmanland             1 0
##   Chapmanmouth            0 1
##   Charlenetown            0 1
##   Charlesbury             1 0
##   Charlesport             0 1
##   Charlottefort           0 1
##   Chaseshire              0 1
##   Chrismouth              0 1
##   Christinehaven          0 1
##   Christinetown           0 1
##   Christopherchester      1 0
##   Christopherport         0 1
##   Christopherville        1 0
##   Clarkborough            0 1
##   Claytonside             1 0
##   Clineshire              1 0
##   Codyburgh               0 1
##   Coffeytown              1 0
##   Colebury                0 1
##   Colemanshire            1 0
##   Collinsburgh            1 0
##   Combsstad               0 1
##   Contrerasshire          1 0
##   Costaburgh              0 1
##   Courtneyfort            0 1
##   Coxhaven                1 0
##   Cranemouth              1 0
##   Crawfordfurt            0 1
##   Cunninghamhaven         0 1
##   Curtisport              0 1
##   Curtisview              1 0
##   Cynthiaside             1 0
##   Daisymouth              1 0
##   Danielview              0 1
##   Davidmouth              0 1
##   Davidside               0 1
##   Davidstad               0 1
```

```
##   Davidton                1 0
##   Davidview               0 1
##   Daviesborough           1 0
##   Davieshaven             1 0
##   Davilachester           0 1
##   Davisfurt               0 1
##   Dayton                  1 0
##   Deannaville             1 0
##   Debraburgh              0 1
##   Derrickhaven            1 0
##   Destinyfurt             0 1
##   Dianashire              1 0
##   Dianaville              0 1
##   Donaldshire             1 0
##   Douglasview             1 0
##   Duffystad               0 1
##   Dustinborough           1 0
##   Dustinchester           1 0
##   Dustinmouth             0 1
##   East Aaron              1 0
##   East Anthony            0 1
##   East Barbara            0 1
##   East Benjaminville      1 0
##   East Breannafurt        0 1
##   East Brettton           0 1
##   East Brianberg          1 0
##   East Brittanyville      0 1
##   East Carlos             1 0
##   East Christopher        1 0
##   East Christopherbury    1 0
##   East Connie             1 0
##   East Dana               0 1
##   East Deborahhaven       1 0
##   East Debraborough       1 0
##   East Donna              0 1
##   East Donnatown          1 0
##   East Eric               0 1
##   East Ericport           0 1
##   East Georgeside         0 1
##   East Graceland          1 0
##   East Heatherside        0 1
##   East Heidi              0 1
##   East Henry              1 0
##   East Jason              0 1
##   East Jennifer           1 0
##   East Jessefort          0 1
##   East John               1 1
##   East Johnport           1 0
##   East Kevinbury          0 1
##   East Lindsey            0 1
```

```
##   East Maureen            0 1
##   East Michaelland        1 0
##   East Michaelmouth       0 1
##   East Michaeltown        1 0
##   East Michele            0 1
##   East Michelleberg       0 1
##   East Mike               0 1
##   East Paul               1 0
##   East Rachaelfurt        0 1
##   East Rachelview         0 1
##   East Ronald             0 1
##   East Samanthashire      0 1
##   East Sharon             0 1
##   East Shawn              0 1
##   East Shawnchester       1 0
##   East Sheriville         1 0
##   East Stephen            0 1
##   East Susanland          1 0
##   East Tammie             0 1
##   East Theresashire       1 0
##   East Tiffanyport        1 0
##   East Timothy            2 0
##   East Timothyport        1 0
##   East Toddfort           1 0
##   East Troyhaven          1 0
##   East Tylershire         0 1
##   East Valerie            1 0
##   East Vincentstad        0 1
##   East Yvonnechester      0 1
##   Edwardmouth             1 0
##   Edwardsmouth            1 0
##   Edwardsport             0 1
##   Elizabethbury           0 1
##   Elizabethmouth          1 0
##   Elizabethport           0 1
##   Elizabethstad           0 1
##   Emilyfurt               1 0
##   Ericksonmouth           0 1
##   Erikville               1 0
##   Erinmouth               1 0
##   Erinton                 0 1
##   Estesfurt               0 1
##   Estradafurt             1 0
##   Estradashire            0 1
##   Evansfurt               1 0
##   Evansville              0 1
##   Faithview               1 0
##   Florestown              0 1
##   Fosterside              0 1
##   Frankbury               0 1
```

```
##   Frankchester            1 0
##   Frankport               0 1
##   Fraziershire            0 1
##   Garciamouth             0 1
##   Garciaside              0 1
##   Garciatown              1 0
##   Garciaview              0 1
##   Garnerberg              1 0
##   Garrettborough          1 0
##   Garychester             1 0
##   Gilbertville            1 0
##   Gomezport               1 0
##   Gonzalezburgh           1 0
##   Grahamberg              0 1
##   Gravesport              1 0
##   Greenechester           1 0
##   Greentown               1 0
##   Greerport               0 1
##   Greerton                1 0
##   Greghaven               1 0
##   Guzmanland              0 1
##   Haleberg                1 0
##   Haleview                1 0
##   Hallfort                1 0
##   Hamiltonfort            0 1
##   Hammondport             1 0
##   Hannahside              1 0
##   Hannaport               0 1
##   Hansenland              0 1
##   Hansenmouth             0 1
##   Harmonhaven             1 0
##   Harperborough           0 1
##   Harrishaven             1 0
##   Harrisonmouth           1 0
##   Hartmanchester          0 1
##   Hartport                1 0
##   Harveyport              0 1
##   Hatfieldshire           1 0
##   Hawkinsbury             0 1
##   Hayesmouth              1 0
##   Heatherberg             0 1
##   Helenborough            0 1
##   Hendrixmouth            0 1
##   Henryfort               1 0
##   Henryland               0 1
##   Hernandezchester        1 0
##   Hernandezfort           1 0
##   Hernandezside           0 1
##   Hernandezville          0 1
##   Hessstad                1 0
```

```
##   Hintonport              0 1
##   Hobbsbury               0 1
##   Holderville             0 1
##   Hollandberg             1 0
##   Hollyfurt               1 0
##   Hubbardmouth            0 1
##   Huffmanchester          0 1
##   Hughesport              0 1
##   Hurleyborough           1 0
##   Ianmouth                1 0
##   Ingramberg              1 0
##   Isaacborough            0 1
##   Jacksonburgh            0 1
##   Jacksonmouth            1 0
##   Jacksonstad             0 1
##   Jacobstad               0 1
##   Jacquelineshire         0 1
##   Jamesberg               1 0
##   Jamesfurt               0 1
##   Jamesmouth              0 1
##   Jamesville              1 0
##   Jamieberg               1 0
##   Jamiefort               1 0
##   Janiceview              1 0
##   Jasminefort             1 0
##   Jayville                1 0
##   Jeffreyburgh            0 1
##   Jeffreymouth            0 1
##   Jeffreyshire            1 0
##   Jenniferhaven           0 1
##   Jenniferstad            1 0
##   Jensenborough           0 1
##   Jensenton               0 1
##   Jeremybury              0 1
##   Jeremyshire             1 0
##   Jessicahaven            0 1
##   Jessicashire            0 1
##   Jessicastad             0 1
##   Joanntown               1 0
##   Joechester              0 1
##   Johnport                1 0
##   Johnsonfort             1 0
##   Johnsontown             0 1
##   Johnsonview             0 1
##   Johnsport               1 0
##   Johnstad                2 0
##   Johnstonmouth           0 1
##   Johnstonshire           1 0
##   Jonathanland            0 1
##   Jonathantown            0 1
```

```
##   Jonesland               1 0
##   Jonesmouth              1 0
##   Jonesshire              0 1
##   Joneston                1 1
##   Jordanmouth             1 0
##   Jordanshire             0 1
##   Jordantown              0 1
##   Josephberg              0 1
##   Josephmouth             0 1
##   Josephstad              0 1
##   Joshuaburgh             1 0
##   Joshuamouth             1 0
##   Juanport                1 0
##   Juliaport               1 0
##   Julietown               0 1
##   Karenmouth              1 0
##   Karenton                1 0
##   Katieport               0 1
##   Kaylashire              1 0
##   Keithtown               0 1
##   Kellytown               1 0
##   Kennedyfurt             1 0
##   Kennethview             1 0
##   Kentmouth               0 1
##   Kevinberg               0 1
##   Kevinchester            1 0
##   Kimberlyhaven           1 0
##   Kimberlymouth           0 1
##   Kimberlytown            1 0
##   Kingchester             0 1
##   Kingshire               1 0
##   Klineside               0 1
##   Knappburgh              1 0
##   Kristineberg            1 0
##   Kristinfurt             0 1
##   Kristintown             0 1
##   Kyleborough             0 1
##   Kylieview               1 0
##   Lake Adrian             1 0
##   Lake Allenville         0 1
##   Lake Amanda             0 1
##   Lake Amy                1 0
##   Lake Angela             1 0
##   Lake Annashire          1 0
##   Lake Beckyburgh         0 1
##   Lake Brandonview        0 1
##   Lake Brian              1 0
##   Lake Cassandraport      0 1
##   Lake Charlottestad      0 1
##   Lake Christopherfurt    0 1
```

```
##   Lake Conniefurt         0 1
##   Lake Courtney           1 0
##   Lake Craigview          0 1
##   Lake Cynthia            1 0
##   Lake Danielle           1 0
##   Lake David              0 2
##   Lake Deannaborough      1 0
##   Lake Deborahburgh       1 0
##   Lake Dustin             0 1
##   Lake Edward             0 1
##   Lake Elizabethside      1 0
##   Lake Evantown           0 1
##   Lake Faith              0 1
##   Lake Gerald             0 1
##   Lake Hailey             1 0
##   Lake Ian                0 1
##   Lake Jacob              1 0
##   Lake Jacqueline         1 0
##   Lake James              0 2
##   Lake Jasonchester       1 0
##   Lake Jennifer           0 1
##   Lake Jenniferton        1 0
##   Lake Jessica            0 1
##   Lake Jessicaville       0 1
##   Lake Jesus              0 1
##   Lake Jillville          1 0
##   Lake John               0 1
##   Lake Johnbury           0 1
##   Lake Jonathanview       1 0
##   Lake Jose               1 1
##   Lake Joseph             1 0
##   Lake Josetown           1 0
##   Lake Joshuafurt         0 1
##   Lake Kevin              1 0
##   Lake Kurtmouth          1 0
##   Lake Lisa               1 0
##   Lake Matthew            0 1
##   Lake Matthewland        1 0
##   Lake Melindamouth       1 0
##   Lake Michael            1 0
##   Lake Michaelport        1 0
##   Lake Michelle           0 1
##   Lake Michellebury       0 1
##   Lake Nicole             1 0
##   Lake Patrick            2 0
##   Lake Rhondaburgh        0 1
##   Lake Stephenborough     0 1
##   Lake Susan              1 1
##   Lake Timothy            1 0
##   Lake Tracy              0 1
```

```
##   Lake Vanessa             0 1
##   Lake Zacharyfurt         1 0
##   Lauraburgh               1 0
##   Laurieside               1 0
##   Lawrenceborough          1 0
##   Lawsonshire              0 1
##   Leahside                 0 1
##   Leonchester              1 0
##   Lesliebury               0 1
##   Lesliefort               1 0
##   Lewismouth               0 1
##   Lindaside                1 0
##   Lindsaymouth             1 0
##   Lisaberg                 1 0
##   Lisafort                 1 0
##   Lisamouth                1 2
##   Lopezberg                0 1
##   Lopezmouth               1 0
##   Loriville                0 1
##   Lovemouth                0 1
##   Luischester              1 0
##   Luisfurt                 1 0
##   Lukeport                 1 0
##   Mackenziemouth           1 0
##   Marcushaven              1 0
##   Mariahview               0 1
##   Mariebury                1 0
##   Mariemouth               1 0
##   Markhaven                0 1
##   Masonhaven               1 0
##   Masseyshire              0 1
##   Mataberg                 1 0
##   Matthewtown              0 1
##   Mauricefurt              0 1
##   Mauriceshire             1 0
##   Mcdonaldfort             1 0
##   Mclaughlinbury           1 0
##   Meaganfort               1 0
##   Meghanchester            0 1
##   Melanieton               0 1
##   Melissachester           0 1
##   Melissafurt              1 0
##   Melissastad              1 0
##   Meyerchester             1 0
##   Meyersstad               0 1
##   Mezaton                  0 1
##   Michaelland              1 0
##   Michaelmouth             1 0
##   Michaelshire             0 1
##   Micheletown              0 1
```

```
##   Michellefort            0 1
##   Michelleside            0 2
##   Millerbury              0 2
##   Millerchester           0 1
##   Millerfort              1 0
##   Millerland              1 0
##   Millerside              0 1
##   Millertown              1 1
##   Millerview              1 0
##   Mollyport               1 0
##   Monicaview              0 1
##   Morganfort              1 0
##   Morganport              0 1
##   Morrismouth             0 1
##   Mosleyburgh             1 0
##   Mullenside              1 0
##   Munozberg               1 0
##   Murphymouth             1 0
##   Nelsonfurt              0 1
##   New Amanda              0 1
##   New Angelview           0 1
##   New Brandy              1 0
##   New Brendafurt          0 1
##   New Charleschester      0 1
##   New Christinatown       0 1
##   New Cynthia             1 0
##   New Daniellefort        0 1
##   New Darlene             0 1
##   New Dawnland            1 0
##   New Debbiestad          0 1
##   New Denisebury          0 1
##   New Frankshire          1 0
##   New Gabriel             1 0
##   New Henry               0 1
##   New Hollyberg           0 1
##   New James               0 1
##   New Jamestown           1 0
##   New Jasmine             1 0
##   New Jay                 0 1
##   New Jeffreychester      1 0
##   New Jessicaport         2 0
##   New Johnberg            1 0
##   New Joshuaport          0 1
##   New Juan                1 0
##   New Julianberg          0 1
##   New Julie               1 0
##   New Karenberg           0 1
##   New Kayla               1 0
##   New Keithburgh          0 1
##   New Lindaberg           0 1
```

```
##   New Lucasburgh          0 1
##   New Marcusbury          0 1
##   New Maria               1 0
##   New Matthew             0 1
##   New Michael             0 1
##   New Michaeltown         1 0
##   New Nancy               0 1
##   New Nathan              1 0
##   New Patriciashire       1 0
##   New Patrick             0 1
##   New Paul                1 0
##   New Rachel              0 1
##   New Rebecca             0 1
##   New Sabrina             0 1
##   New Sean                1 0
##   New Shane               1 0
##   New Sharon              1 0
##   New Sheila              2 0
##   New Sonialand           1 0
##   New Steve               1 0
##   New Tammy               0 1
##   New Taylorburgh         1 0
##   New Teresa              0 1
##   New Theresa             0 1
##   New Thomas              0 1
##   New Timothy             0 1
##   New Tina                0 1
##   New Tinamouth           1 0
##   New Traceystad          1 0
##   New Travis              1 0
##   New Travistown          0 1
##   New Tyler               1 0
##   New Wanda               1 0
##   New Williammouth        0 1
##   New Williamville        0 1
##   Newmanberg              1 0
##   Nicholasland            0 1
##   Nicholasport            1 0
##   North Aaronburgh        0 1
##   North Aaronchester      0 1
##   North Alexandra         1 0
##   North Anaport           1 0
##   North Andrew            0 1
##   North Andrewstad        0 1
##   North Angelastad        0 1
##   North Angelatown        0 1
##   North Anna              1 0
##   North April             0 1
##   North Brandon           1 0
##   North Brittanyburgh     0 1
```

```
##   North Cassie             0 1
##   North Charlesbury        0 1
##   North Christopher        1 0
##   North Daniel             1 1
##   North Debra              1 0
##   North Debrashire         0 1
##   North Derekville         0 1
##   North Destiny            0 1
##   North Elizabeth          1 0
##   North Frankstad          1 0
##   North Garyhaven          1 0
##   North Isabellaville      1 0
##   North Jenniferburgh      0 1
##   North Jeremyport         1 0
##   North Jessicaville       0 1
##   North Johnside           1 0
##   North Johntown           0 1
##   North Jonathan           0 1
##   North Joshua             1 0
##   North Katie              0 1
##   North Kennethside        1 0
##   North Kevinside          0 1
##   North Kimberly           0 1
##   North Kristine           1 0
##   North Lauraland          0 1
##   North Laurenview         1 0
##   North Leonmouth          1 0
##   North Lisachester        1 0
##   North Loriburgh          1 0
##   North Mark               0 1
##   North Maryland           0 1
##   North Mercedes           0 1
##   North Michael            0 1
##   North Monicaville        1 0
##   North Randy              1 0
##   North Raymond            1 0
##   North Regina             0 1
##   North Ricardotown        0 1
##   North Richardburgh       0 1
##   North Ronaldshire        1 0
##   North Russellborough     0 1
##   North Samantha           0 1
##   North Sarashire          0 1
##   North Shannon            1 0
##   North Stephanieberg      1 0
##   North Tara               1 0
##   North Tiffany            1 0
##   North Tracyport          1 0
##   North Tylerland          1 0
##   North Virginia           0 1
```

```
##   North Wesleychester     1 0
##   Novaktown               1 0
##   Odomville               1 0
##   Olsonside               0 1
##   Olsonstad               0 1
##   Palmerside              0 1
##   Pamelamouth             2 0
##   Parkerhaven             1 0
##   Patriciahaven           1 0
##   Patrickmouth            1 0
##   Pattymouth              0 1
##   Paulhaven               1 0
##   Paulport                1 0
##   Paulshire               1 0
##   Pearsonfort             1 0
##   Penatown                0 1
##   Perezland               1 0
##   Perryburgh              0 1
##   Petersonfurt            0 1
##   Phelpschester           1 0
##   Philipberg              0 1
##   Phillipsbury            0 1
##   Port Aliciabury         1 0
##   Port Angelamouth        0 1
##   Port Anthony            1 0
##   Port Aprilville         0 1
##   Port Beth               0 1
##   Port Blake              0 1
##   Port Brenda             0 1
##   Port Brian              0 1
##   Port Brianfort          1 0
##   Port Brittanyville      1 0
##   Port Brookeland         0 1
##   Port Calvintown         1 0
##   Port Cassie             0 1
##   Port Chasemouth         1 0
##   Port Christina          0 1
##   Port Christinemouth     1 0
##   Port Christopher        0 1
##   Port Christopherborough 0 1
##   Port Crystal            0 1
##   Port Daniel             1 0
##   Port Danielleberg       1 0
##   Port Davidland          1 0
##   Port Dennis             0 1
##   Port Derekberg          0 1
##   Port Destiny            1 0
##   Port Douglasborough     0 1
##   Port Elijah             1 0
##   Port Eric               0 1
```

```
##   Port Erikhaven          0 1
##   Port Erinberg           0 1
##   Port Eugeneport         1 0
##   Port Georgebury         0 1
##   Port Gregory            1 0
##   Port Jacqueline         1 0
##   Port Jacquelinestad     1 0
##   Port James              1 0
##   Port Jasmine            1 0
##   Port Jason              1 1
##   Port Jefferybury        0 1
##   Port Jeffrey            1 0
##   Port Jennifer           0 1
##   Port Jessica            0 1
##   Port Jessicamouth       1 0
##   Port Jodi               1 0
##   Port Joshuafort         0 1
##   Port Juan               1 1
##   Port Julie              1 1
##   Port Karenfurt          1 0
##   Port Katelynview        0 1
##   Port Kathleenfort       0 1
##   Port Kevinborough       1 0
##   Port Lawrence           0 1
##   Port Maria              1 0
##   Port Mathew             1 0
##   Port Melissaberg        0 1
##   Port Melissastad        1 0
##   Port Michaelmouth       0 1
##   Port Michealburgh       0 1
##   Port Mitchell           0 1
##   Port Patrickton         0 1
##   Port Paultown           0 1
##   Port Rachel             0 1
##   Port Raymondfort        1 0
##   Port Robin              1 0
##   Port Sarahhaven         0 1
##   Port Sarahshire         0 1
##   Port Sherrystad         0 1
##   Port Stacey             1 0
##   Port Stacy              1 0
##   Port Susan              1 0
##   Port Whitneyhaven       1 0
##   Portermouth             1 0
##   Pottermouth             0 1
##   Princebury              1 0
##   Pruittmouth             1 0
##   Rachelhaven             1 0
##   Ramirezhaven            0 1
##   Ramirezland             1 0
```

```
##   Ramirezside             0 1
##   Ramirezton              1 0
##   Ramosstad               1 0
##   Randolphport            1 0
##   Randyshire              1 0
##   Rebeccamouth            0 1
##   Reginamouth             0 1
##   Reneechester            0 1
##   Reyesfurt               1 0
##   Reyesland               1 0
##   Rhondaborough           1 0
##   Richardshire            0 1
##   Richardsland            1 0
##   Richardsonland          0 1
##   Richardsonmouth         1 0
##   Richardsonshire         0 1
##   Richardsontown          1 0
##   Rickymouth              1 0
##   Riggsstad               1 0
##   Rivasland               0 1
##   Robertbury              1 0
##   Robertfurt              0 2
##   Robertmouth             1 0
##   Robertside              0 1
##   Robertsonburgh          0 1
##   Robertstown             0 1
##   Roberttown              0 1
##   Robinsonland            1 0
##   Robinsontown            0 1
##   Rochabury               0 1
##   Rogerburgh              0 1
##   Rogerland               1 0
##   Ronaldport              0 1
##   Ronniemouth             0 1
##   Russellville            0 1
##   Ryanhaven               0 1
##   Sabrinaview             1 0
##   Salazarbury             0 1
##   Samanthaland            0 1
##   Samuelborough           1 0
##   Sanchezland             1 0
##   Sanchezmouth            1 0
##   Sandersland             1 0
##   Sanderstown             0 1
##   Sandraland              1 0
##   Sandrashire             0 1
##   Sandraville             1 0
##   Sarafurt                1 0
##   Sarahland               0 1
##   Sarahton                1 0
```

```
##   Sellerstown             1 0
##   Shaneland               1 0
##   Sharpberg               1 0
##   Shawnside               1 0
##   Shawstad                1 0
##   Shelbyport              1 1
##   Sherrishire             1 0
##   Shirleyfort             1 0
##   Silvaton                0 1
##   Smithburgh              1 0
##   Smithside               0 1
##   Smithtown               1 0
##   South Aaron             0 1
##   South Adam              0 1
##   South Adamhaven         1 0
##   South Alexisborough     0 1
##   South Blakestad         1 0
##   South Brian             1 0
##   South Cathyfurt         0 1
##   South Christopher       1 0
##   South Corey             1 0
##   South Cynthiashire      0 1
##   South Daniel            0 1
##   South Daniellefort      1 0
##   South Davidhaven        0 1
##   South Davidmouth        0 1
##   South Denise            1 0
##   South Denisefurt        1 0
##   South Dianeshire        1 0
##   South George            0 1
##   South Henry             0 1
##   South Jackieberg        0 1
##   South Jade              0 1
##   South Jaimeview         1 0
##   South Jasminebury       0 1
##   South Jeanneport        0 1
##   South Jennifer          1 0
##   South Jessica           0 1
##   South John              0 1
##   South Johnnymouth       0 1
##   South Kyle              0 1
##   South Lauraton          0 1
##   South Lauratown         0 1
##   South Lisa              0 2
##   South Manuel            1 0
##   South Margaret          0 1
##   South Mark              0 1
##   South Meghan            0 1
##   South Meredithmouth     1 0
##   South Pamela            1 0
```

```
##   South Patrickfort       1 0
##   South Peter             0 1
##   South Rebecca           0 1
##   South Renee             1 0
##   South Robert            1 0
##   South Ronald            1 0
##   South Stephanieport     1 0
##   South Tiffanyton        0 1
##   South Tomside           1 0
##   South Troy              1 0
##   South Vincentchester    0 1
##   South Walter            0 1
##   Staceyfort              0 1
##   Stephenborough          1 0
##   Stewartbury             1 0
##   Suzannetown             0 1
##   Sylviaview              1 0
##   Tammymouth              0 1
##   Tammyshire              0 1
##   Taylorberg              1 0
##   Taylorhaven             0 1
##   Taylormouth             0 1
##   Taylorport              1 0
##   Teresahaven             1 0
##   Thomasstad              1 0
##   Thomasview              1 0
##   Timothyfurt             0 1
##   Timothymouth            0 1
##   Timothyport             0 1
##   Timothytown             1 0
##   Tinachester             1 0
##   Tinaton                 0 1
##   Townsendfurt            1 0
##   Tracyhaven              0 1
##   Tranland                1 0
##   Troyville               1 0
##   Turnerchester           0 1
##   Turnerview              1 0
##   Turnerville             1 0
##   Tylerport               0 1
##   Valerieland             1 0
##   Vanessastad             0 1
##   Vanessaview             0 1
##   Villanuevastad          1 0
##   Villanuevaton           1 0
##   Wademouth               1 0
##   Wadestad                1 0
##   Wagnerchester           1 0
##   Wallacechester          1 0
##   Walshhaven              1 0
```

```
##   Waltertown               0 1
##   Watsonfort               1 0
##   Welchshire               0 1
##   Wendyton                 1 0
##   Wendyville               0 1
##   West Alice               1 0
##   West Alyssa              1 0
##   West Amanda              0 2
##   West Andrew              1 0
##   West Angela              1 0
##   West Angelabury          1 0
##   West Annefort            0 1
##   West Aprilport           0 1
##   West Arielstad           1 0
##   West Barbara             1 0
##   West Benjamin            1 0
##   West Brad                0 1
##   West Brandonton          0 1
##   West Brenda              1 0
##   West Carmenfurt          1 0
##   West Casey               0 1
##   West Chloeborough        0 1
##   West Christopher         0 1
##   West Colin               1 0
##   West Connor              0 1
##   West Courtney            1 0
##   West Daleborough         1 0
##   West Dannyberg           1 0
##   West David               0 1
##   West Dennis              1 0
##   West Derekmouth          0 1
##   West Dylanberg           0 1
##   West Eduardotown         0 1
##   West Ericaport           0 1
##   West Ericfurt            0 1
##   West Gabriellamouth      0 1
##   West Gregburgh           1 0
##   West Guybury             1 0
##   West James               0 1
##   West Jane                0 1
##   West Jeremyside          0 1
##   West Jessicahaven        0 1
##   West Jodi                1 0
##   West Joseph              1 0
##   West Julia               0 1
##   West Justin              0 1
##   West Katiefurt           0 1
##   West Kevinfurt           0 1
##   West Lacey               1 0
##   West Leahton             0 1
```

```
##   West Lindseybury        0 1
##   West Lisa               1 0
##   West Lucas              1 0
##   West Mariafort          1 0
##   West Melaniefurt        0 1
##   West Melissashire       0 1
##   West Michaelhaven       1 0
##   West Michaelport        1 0
##   West Michaelshire       1 0
##   West Michaelstad        1 0
##   West Pamela             0 1
##   West Randy              0 1
##   West Raymondmouth       0 1
##   West Rhondamouth        1 0
##   West Ricardo            0 1
##   West Richard            0 1
##   West Robertside         1 0
##   West Roytown            1 0
##   West Russell            1 0
##   West Ryan               0 1
##   West Samantha           1 0
##   West Shannon            0 2
##   West Sharon             1 0
##   West Shaun              1 0
##   West Steven             2 0
##   West Sydney             1 0
##   West Tanner             1 0
##   West Tanya              0 1
##   West Terrifurt          1 0
##   West Thomas             1 0
##   West Tinashire          0 1
##   West Travismouth        0 1
##   West Wendyland          1 0
##   West William            0 1
##   West Zacharyborough     1 0
##   Westshire               0 1
##   Whiteport               0 1
##   Whitneyfort             1 0
##   Wilcoxport              0 1
##   Williammouth            0 1
##   Williamport             1 0
##   Williamsborough         0 1
##   Williamsfort            0 1
##   Williamsmouth           0 1
##   Williamsport            1 2
##   Williamsside            1 0
##   Williamstad             0 1
##   Wilsonburgh             1 0
##   Wintersfort             1 0
##   Wongland                1 0
```

```
##   Wrightburgh              2 0
##   Wrightview               0 1
##   Yangside                 0 1
##   Youngburgh               1 0
##   Youngfort                0 1
##   Yuton                    0 1
##   Zacharystad              1 0
##   Zacharyton               0 1
```

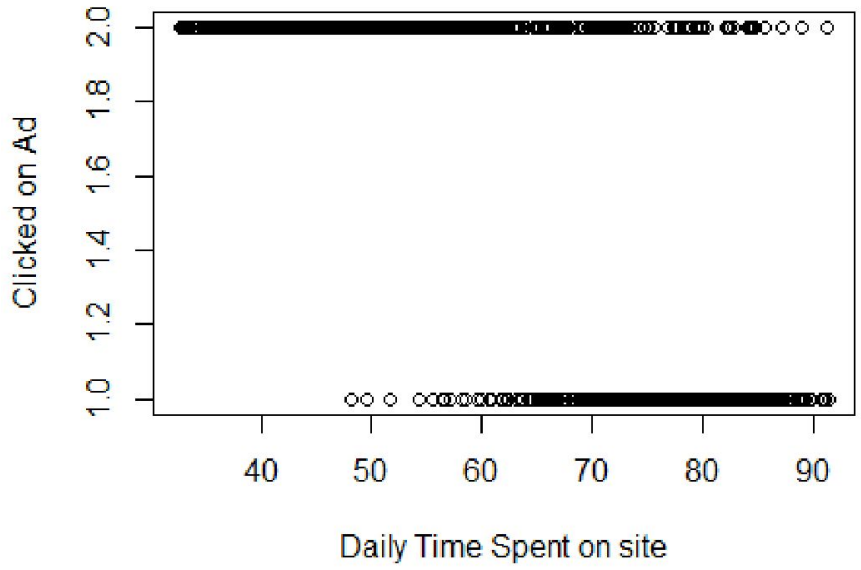
Most cities have at least 1 or 0 clicks on ads. Only a few cities such as Lake David, Lake James, Lisamouth have 2 clicks on ads.

Scatterplots

For continuos numerical columns, we will make scatter plots to establish the relationships between the variables.

Daily time spent versus ads being clicked

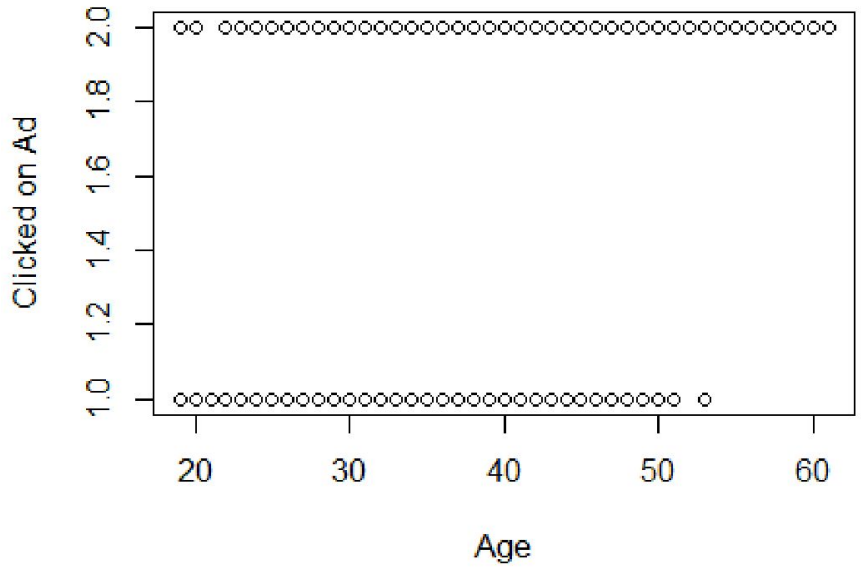
```
# scatter plot of daily time spent versus ad being clicked
plot(data$daily_time_spent_on_site, data$clicked_on_ad, ylab = "Clicked on
Ad", xlab = "Daily Time Spent on site")
```



Users that clicked on the ads are clustered between time 2-65minutes, thereafter, the scatter begins to get dispersed. Users who spent more time on the site did not click on the ad.

Age versus ad being clicked

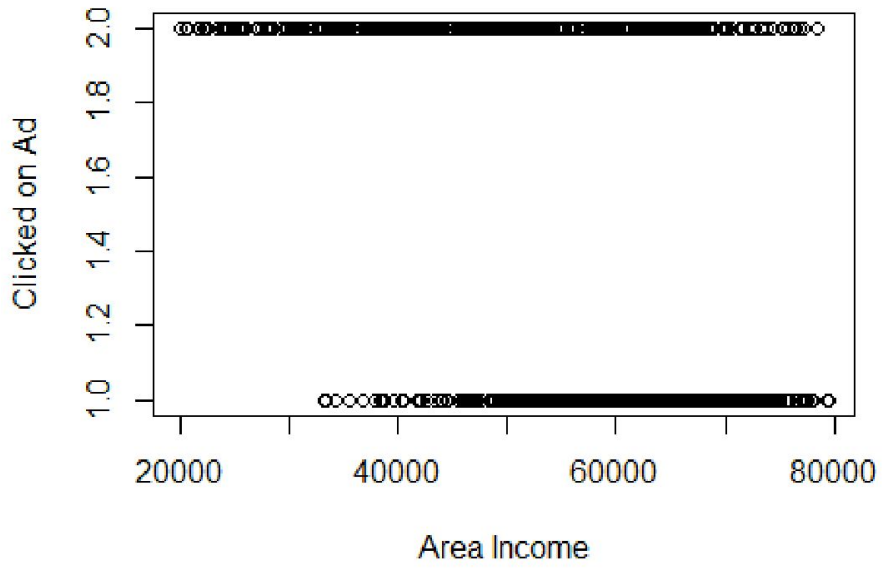
```
# age versus ad being clicked
plot(data$age, data$clicked_on_ad, ylab = "Clicked on Ad", xlab = "Age")
```



The age of a user is not significant to determining whether they will click an ad or not since all users from ages 18 to 60 clicked on the ad. It is notable that older users did click on the ads. This includes ages from 54 and above. It could be because they have all the time to do so as most of them are probably retired.

Area income versus ad being clicked

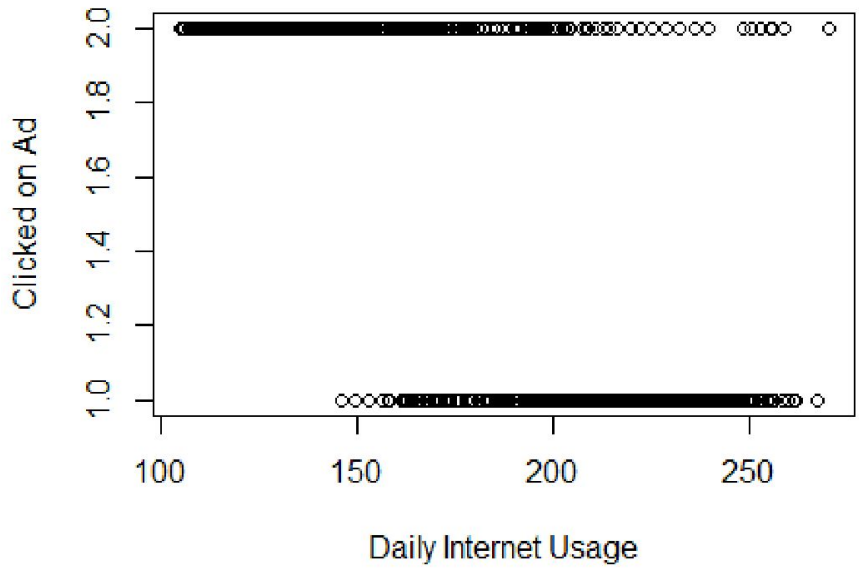
```
# area income versus ad being clicked
plot(data$area_income, data$clicked_on_ad, ylab = "Clicked on Ad", xlab =
"Area Income")
```



All users visiting the site and with a low area income clicked on the ads. This includes users with income below 33000.

Daily Internet Usage versus ads being clicked

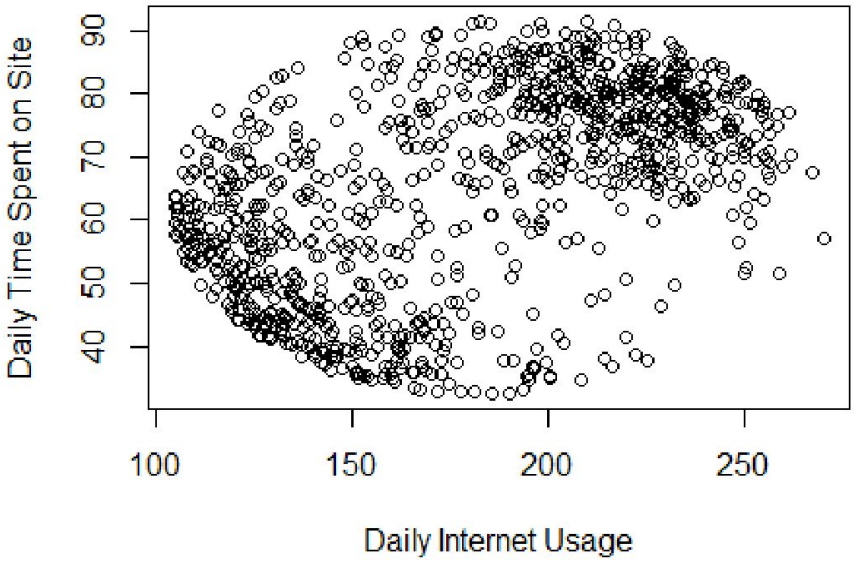
```
# daily internet usage versus ad being clicked
plot(data$daily_internet_usage, data$clicked_on_ad, ylab = "Clicked on Ad",
xlab = "Daily Internet Usage")
```



Users with a low daily internet usage clicked on the ads. These are users with daily internet usage below 150mbs of data. We expect a similar trend on the daily internet usage and daily time spent on the site. We can make a plot to see this relationship.

Daily Internet Usage versus daily time spent on the site

```
# daily internet usage versus daily time spent on site
plot(data$daily_internet_usage, data$daily_time_spent_on_site, ylab = "Daily
Time Spent on Site", xlab = "Daily Internet Usage")
```



There are clusters concentrated at the low left and upper right of the plot. Most users with a low data bundle usage per day spent less time on the site while users who spent more time on the site had more daily data bundles to use.

Correlation Matrix

Find the correlations of the numerical columns and make a correlation matrix plot

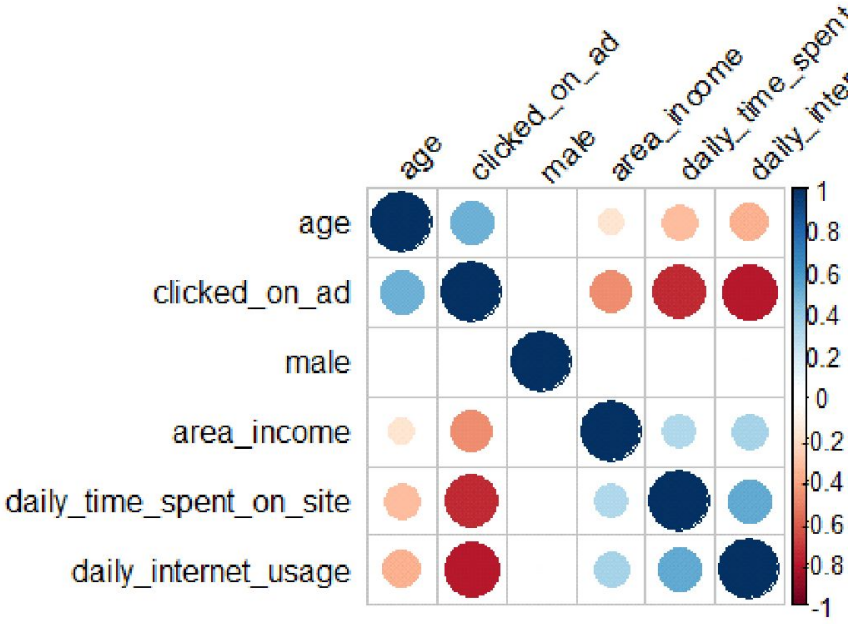
```
# find the correlations and round them off to 2 decimal places
res <- round(cor(numericals), 2)
# round(res, 2)
res
```

```
##                           daily_time_spent_on_site   age area_income
## daily_time_spent_on_site                     1.00 -0.33        0.31
## age                                         -0.33  1.00       -0.18
## area_income                                  0.31 -0.18        1.00
## daily_internet_usage                         0.52 -0.37        0.34
## male                                        -0.02 -0.02        0.00
## clicked_on_ad                               -0.75  0.49       -0.48
##                           daily_internet_usage  male clicked_on_ad
## daily_time_spent_on_site                 0.52 -0.02         -0.75
## age                                     -0.37 -0.02          0.49
## area_income                              0.34  0.00         -0.48
## daily_internet_usage                     1.00  0.03         -0.79
## male                                     0.03  1.00         -0.04
## clicked_on_ad                           -0.79 -0.04          1.00
```

```
library(corrplot)

## corrplot 0.84 loaded

corrplot(res, type = "full", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



- There is positive correlation between daily time spent on site and daily internet usage. This is accustomed to the fact that more data usage equals to more time spent on the internet which equals to the time spent on the site by a user.
- There is a negative correlation between daily time spent on site, daily internet used and whether a user clicked on ad.
- There is a slight correlation of 0.49 between age and whether or not a user clicked on ad.
- There is a slight negative correlation of -0.48 between area income and whether or not a user clicked on ad.
- The gender column does not exhibit strong or noticable relationships with the other variables.

9. Implementing the Solution

Modeling

Selecting our features

We will select only numerical variables from the data to use for modelling

```
new_data <- data[, c(1,2,3,4,7,10,11,12,13)]
colnames(new_data)
```

```
## [1] "daily_time_spent_on_site" "age"
## [3] "area_income"              "daily_internet_usage"
## [5] "male"                     "month"
## [7] "day"                      "hour"
## [9] "clicked_on_ad"
```

Then convert categorical features that are factors to numeric variables

```
# make datatype conversions
new_data$male <- as.numeric(new_data$male)
new_data$month <- as.numeric(new_data$month)
new_data$day <- as.numeric(new_data$day)
new_data$hour <- as.numeric(new_data$hour)
new_data$clicked_on_ad <- as.numeric(new_data$clicked_on_ad)

# check the data types
str(new_data)
```

```
## 'data.frame':    1000 obs. of  9 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 ...
##  $ male                    : num  1 2 1 2 1 2 1 2 2 2 ...
##  $ month                   : num  3 4 3 1 6 5 1 3 4 7 ...
##  $ day                     : num  27 4 13 10 3 19 28 7 18 11 ...
##  $ hour                    : num  1 2 21 3 4 15 21 2 10 2 ...
##  $ clicked_on_ad           : num  1 1 1 1 1 1 1 2 1 1 ...
```

Normalizing the data

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
# Creating a random number equal 80% of total number of rows
ran <- sample(1:nrow(new_data),0.8 * nrow(new_data))

# the normalization function is created
```

```r
normalize <- function(x){
  return ((x-min(x)) / (max(x)-min(x)))
}
# Normalization function is applied to the dataframe
newdata_normalized <- as.data.frame(lapply(new_data[,
c(1,2,3,4,5,6,7,8)],normalize))
head(newdata_normalized)
```

```
##   daily_time_spent_on_site       age area_income daily_internet_usage male
## 1                0.6178820 0.3809524   0.7033117            0.9160310    0
## 2                0.8096209 0.2857143   0.8143826            0.5387456    1
## 3                0.6267211 0.1666667   0.6688882            0.7974331    0
## 4                0.7062723 0.2380952   0.5851850            0.8542802    1
## 5                0.6080231 0.3809524   0.9059586            0.7313234    0
## 6                0.4655788 0.0952381   0.6684784            0.7383460    1
##       month        day       hour
## 1 0.3333333 0.86666667 0.00000000
## 2 0.5000000 0.10000000 0.04347826
## 3 0.3333333 0.40000000 0.86956522
## 4 0.0000000 0.30000000 0.08695652
## 5 0.8333333 0.06666667 0.13043478
## 6 0.6666667 0.60000000 0.60869565
```

Splitting data into training and testing sets

```r
# The training dataset extracted
clicked_train <- newdata_normalized[ran,]

# The test dataset extracted
clicked_test <- newdata_normalized[-ran,]

# training target
train_target <- as.factor(new_data[ran,9])

#testing target
test_target <- as.factor(new_data[-ran,9])
```

Modeling with K-Nearest Neighbors

```r
#load libraries and model using knn
library(class)
model_knn <- knn(clicked_train,clicked_test,cl=train_target,k=3)
```

Model evaluation

```r
# Creating the confusion matrix
tb <- table(model_knn,test_target)
tb
```

```
##          test_target
## model_knn   1   2
```

```
##          1 107   7
##          2   2  84
```

```
# Checking the accuracy
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
accuracy(tb)
```

```
## [1] 95.5
```

Our model has performed quite well with an accuracy of 95.5% and a few misclassification errors in the confusion matrix.

Challenging the solution

We can model using SVM to see how the model will perform and compare the results with the KNN model.

Split data into training and test sets(80:20 ratio)

```
intrain <- createDataPartition(y = new_data$clicked_on_ad, p= 0.8, list = FALSE)
training <- new_data[intrain,]
testing <- new_data[-intrain,]
```

The next step here is to build a suitable SVM model for the predicting whether a user visiting the site will click on the ad.

Modelling using SVM

```
# load libraries
library(rpart)
library(kernlab)
```

```
##
## Attaching package: 'kernlab'
```

```
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```
# convert outcome/target variable to factor so that we perfom classification
training$clicked_on_ad <- as.factor(training$clicked_on_ad)


#modeling
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

svm_Linear <- train(clicked_on_ad ~ .,
                    data = training,
                    method = "svmLinear",
                    trControl=trctrl,
```

```
                    preProcess = c("center", "scale"),
                    tuneLength = 10)

# check the results of the SVM model
svm_Linear
```

```
## Support Vector Machines with Linear Kernel
##
## 800 samples
##   8 predictor
##   2 classes: '1', '2'
##
## Pre-processing: centered (8), scaled (8)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 720, 720, 720, 720, 720, 720, ...
## Resampling results:
##
##   Accuracy  Kappa
##   0.97      0.94
##
## Tuning parameter 'C' was held constant at a value of 1
```

Model Evaluation

```
# predicting
test_pred <- predict(svm_Linear, newdata = testing)

# print confusion matrix
confusionMatrix(table(test_pred, testing$clicked_on_ad))
```

```
## Confusion Matrix and Statistics
##
##
## test_pred  1  2
##         1 96  4
##         2  4 96
##
##                Accuracy : 0.96
##                  95% CI : (0.9227, 0.9826)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.92
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.96
##             Specificity : 0.96
##          Pos Pred Value : 0.96
##          Neg Pred Value : 0.96
##              Prevalence : 0.50
```

```
##          Detection Rate : 0.48
##    Detection Prevalence : 0.50
##       Balanced Accuracy : 0.96
##
##        'Positive' Class : 1
##
```

The accuracy of the model is 96% and with a few misclassification errors in the confusion matrix.

This value is a higher than the KNN model but only by a small percentage. We can therefore conclude that the SVM model is better than the KNN model.

10. Conclusion

The results obtained from the EDA process will be used to make conclusions:

- The dataset was already slightly biased on the gender. There were more women than men visiting the site hence it more females than males clicked on the ads.
- Users who spent less time online were more likely to click on the ad than people who spent more time. As observed, these users also have a low daily internet usage.
- People with lower area incomes clicked more on the ad than people with higher area incomes.
- The month of February and the 3rd days of the month were prime times for ad clicking. For the 31st days and the month of July, not so much.
- Prime times for ad clicking is at 9am in the morning but this gets lower as it gets to 10am which registered low number of ad clicks.

11. Recommendations

The target audience for the enterpreneur is:

- Users with low income
- Users who spend low on daily internet

The target time for advertising the course and displaying ads is at 9am.

The entrepreneur can customize her ads in a way that she gets the attention of users visiting the site in the morning. She can also customize her ads to attract more users including those with a higher income.

She can customize her ads on the online cryptography course by reducing the price. It could be that few users are clicking on the ad because the course is highly priced. Low

priced(affordable) products are relatively attractive to more users, which could mean more traffic to the site.

Use the SVM model to predict whether a site visitor will click on the ad or not since it performs better than the KNN model.