```r
#1 Get Working Directory Path to save the file
getwd()
```

```
## [1] "C:/Users/Nupur Shrinet/Documents"
```

```r
#2 Reading the csv "ecommerce-data.csv" into a data.frame
ecomm.df<-read.csv("ecommerce-data.csv")
```

```r
#3 Opening the dataframe and reviewing the column labels and type
View(ecomm.df)
```

```r
#4 Learning about the structure of the data
str(ecomm.df)
```

```
## 'data.frame':    1593 obs. of  45 variables:
##  $ dateTime                    : Factor w/ 1558 levels "7/25/2014 14:10",..: 1 2 3 4 5 6 14 7 8 9 ...
##  $ country                     : Factor w/ 44 levels "Australia","Barbados",..: 44 44 44 44 44 44 17 4
##  $ city                        : Factor w/ 980 levels "","Abilene","Abingdon",..: 563 25 76 158 132 4
##  $ region                      : Factor w/ 110 levels "","0","1","10",..: 67 94 102 104 35 40 10 80 1
##  $ screenRed                   : Factor w/ 91 levels "1012x569","1024x552",..: 29 34 76 35 17 43 76 7
##  $ surveyType                  : Factor w/ 3 levels "At Arrival and Exit",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ purposeProductInfo          : Factor w/ 2 levels "","Products": 2 1 1 2 1 2 1 2 1 1 ...
##  $ purposeBuyFromSite          : Factor w/ 2 levels "","Buy from this site": 1 2 1 1 1 1 1 1 1 2 ...
##  $ purposeComparePricing       : Factor w/ 2 levels "","Compare pricing": 1 2 2 1 1 2 1 1 1 1 ...
##  $ purposeInfoAndResources     : Factor w/ 2 levels "","Resources": 2 1 1 1 2 1 2 1 1 1 ...
##  $ purposeInfoOnOrder          : Factor w/ 2 levels "","Order info": 1 1 1 1 1 1 1 1 1 1 ...
##  $ purposeOther                : Factor w/ 2 levels "","Other": 1 1 1 1 1 1 1 1 2 1 ...
##  $ taskFindWhatLookingFor      : Factor w/ 4 levels "","Most or all of it",..: 1 1 1 2 2 2 4 2 4 2 ..
##  $ concernShippingCost         : Factor w/ 2 levels "","Shipping costs": 1 1 1 2 1 1 1 1 1 1 ...
##  $ concernDeliverySpeed        : Factor w/ 2 levels "","Fast delivery": 1 1 1 1 1 1 1 1 1 1 ...
##  $ concernWarranties           : Factor w/ 2 levels "","Warranties/product guarantees": 1 1 1 1 1 1 1
##  $ concernEaseToReturnProduct  : Factor w/ 2 levels "","Ease of returning (if I am not satisfied with
##  $ concernProductSafety        : Factor w/ 2 levels "","Product safety": 1 1 1 1 1 1 1 1 1 1 ...
##  $ concernRightForMyChild      : Factor w/ 2 levels "","Whether this is right for my child": 1 1 1 1 1
##  $ concernProductQuality       : Factor w/ 2 levels "","Product durability/quality": 2 1 1 1 1 1 1 1 1
##  $ concernProductEffectiveness: Factor w/ 2 levels "","Product effectiveness/will it work": 2 1 1 1 1
##  $ concernOther                : Factor w/ 2 levels "","Other": 1 1 1 1 1 1 1 1 1 1 ...
##  $ concernNone                 : Factor w/ 2 levels "","None / no uncertainties": 1 1 1 1 1 1 1 1 2 1
##  $ intentWasPlanningToBuy      : Factor w/ 4 levels "","No","Partially (I was considering it)",..: 1 4
##  $ profile                     : Factor w/ 8 levels "0","Friend/family friend",..: 5 5 5 6 8 6 3 4 8 5
##  $ whenSiteUsed                : Factor w/ 6 levels "","In the past month",..: 3 4 6 6 6 6 3 3 6 6 ..
##  $ purchasedBefore             : Factor w/ 4 levels "","No","Yes, more than once",..: 4 4 1 1 1 1 2 4
##  $ purchasedWhen               : Factor w/ 5 levels "","In the past month",..: 2 4 1 1 1 1 1 2 1 1 ..
##  $ productKnewWhatWanted       : Factor w/ 4 levels "","No","Somewhat",..: 4 4 4 3 1 3 1 1 1 4 ...
##  $ productSiteHasWhatWanted    : Factor w/ 5 levels "","No","Not sure",..: 1 1 1 5 1 5 1 1 1 5 ...
##  $ purchaseExpectInNextMonth   : int  5 3 3 3 5 3 5 NA 5 4 ...
##  $ siteFirstHeardAbout         : Factor w/ 6 levels "","In the past hour",..: 4 6 5 2 5 2 3 1 5 1 ...
##  $ age                         : Factor w/ 9 levels "","18-24","25-34",..: 3 4 4 3 6 2 6 1 5 1 ...
##  $ gender                      : Factor w/ 4 levels "","Female","Male",..: 2 2 2 2 2 4 2 1 2 1 ...
##  $ behavNumVisits              : int  13 3 2 1 1 1 4 1 2 2 ...
##  $ behavReferral               : Factor w/ 9 levels "","Branded Search",..: 3 9 9 9 6 8 3 9 6 9 ...
##  $ behavPageviews              : Factor w/ 6 levels "0","1","10+",..: 5 2 3 3 2 3 3 5 3 6 ...
```

```
##  $ behavHomePage             : int  1 0 0 0 0 1 0 1 1 1 ...
##  $ behavDetailProdA          : int  1 0 0 1 0 1 1 0 1 1 ...
##  $ behavDetailProdB          : int  0 0 0 1 0 1 1 1 1 0 ...
##  $ behavDetailProdC          : int  0 0 0 0 0 0 1 0 1 0 ...
##  $ behavAnySolution          : int  0 0 1 1 0 0 1 0 1 0 ...
##  $ behavAnySale              : int  0 0 1 0 0 0 1 0 1 1 ...
##  $ behavCart                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ behavConversion           : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
# Question 1
# Answer 1 - Using str() we can get the number of observation i.e. 1593 and number of variables i.e. 45
```

```
# Question 2
# Creating table with country (factor) to get the frequency of the visits
country_of_origin<-table(ecomm.df$country)
# Sort the table in decreasing order of frequency to get the most visits by country
sort(country_of_origin,decreasing = TRUE)
```

```
##
##        United States               Canada            Australia
##                 1361                   62                   50
##       United Kingdom                India         South Africa
##                   31                   13                    8
##          Puerto Rico               Israel          Netherlands
##                    6                    4                    4
## United Arab Emirates               Brazil           Costa Rica
##                    4                    3                    3
##              Denmark              Ireland             Malaysia
##                    3                    3                    3
##               Mexico              Germany                Malta
##                    3                    2                    2
##              Nigeria          Philippines             Barbados
##                    2                    2                    1
##             Botswana             Colombia               France
##                    1                    1                    1
##                Haiti                Italy                Japan
##                    1                    1                    1
##               Kuwait              Namibia          New Zealand
##                    1                    1                    1
##               Norway               Panama               Poland
##                    1                    1                    1
##             Portugal              Romania         Saudi Arabia
##                    1                    1                    1
##            Singapore             Slovenia                Spain
##                    1                    1                    1
##               Sweden             Thailand  Trinidad and Tobago
##                    1                    1                    1
##               Turkey              Ukraine
##                    1                    1
```

```
# Answer 2 - After United States (1362), Canada has the most site vists with 62 visits
```

```r
# Question 3
# Creating a two-way table with intentPlanningtoBuy broken out by profile
planningtobuy_profile<-table(ecomm.df$intentWasPlanningToBuy,ecomm.df$profile)
View(planningtobuy_profile)
```

```r
# Question 4
# Creating propotion for each profile using "margin=2", to get 100 % for each profile.
prop.table((planningtobuy_profile),margin=2)
```

```
##
##                                          0 Friend/family friend
##                                 1.00000000           0.95652174
##   No                            0.00000000           0.00000000
##   Partially (I was considering it) 0.00000000           0.04347826
##   Yes                           0.00000000           0.00000000
##
##                                 Health Professional      Other     Parent
##                                           0.73285199 0.71641791 0.63144330
##   No                                      0.04332130 0.03731343 0.02061856
##   Partially (I was considering it)        0.13718412 0.14179104 0.22422680
##   Yes                                     0.08664260 0.10447761 0.12371134
##
##                                 Person with [condition A]   Relative
##                                                0.76923077 0.72897196
##   No                                           0.00000000 0.03738318
##   Partially (I was considering it)             0.15384615 0.13084112
##   Yes                                          0.07692308 0.10280374
##
##                                    Teacher
##                                 0.73991031
##   No                            0.04035874
##   Partially (I was considering it) 0.15695067
##   Yes                           0.06278027
```

```r
# Propotion of parents intented to buy include responses "Partially" and "Yes" i.e. 22.4% & 12.4% respe

# Propotion of teacher who did buy include responses "Yes" only i.e. 6.2%
```

```r
# Question 5
# Subsetting the dataframe to create another dataframe for only United States records
ecomm_USA.df<-subset(ecomm.df, ecomm.df$country=="United States")
# creating a freq table for regions in USA
region_of_origin<-table(ecomm_USA.df$region)
# sorting the freq table in descending order of number of site visits
sorted_region_of_origin<-sort(region_of_origin, decreasing = TRUE)
# Finding the region with maximum site visits

head(sorted_region_of_origin)
```

```
##
## TX NY CA IL PA FL
## 94 92 90 64 57 56
```

```r
max(sorted_region_of_origin)
```

```
## [1] 94
```

```r
# Answer 5 - Texas (TX) has the most site visits with 94 vistis compared to other states /regions
```

```r
# Question 6
# Using which.max() to get to the above result
which.max(sorted_region_of_origin)
```

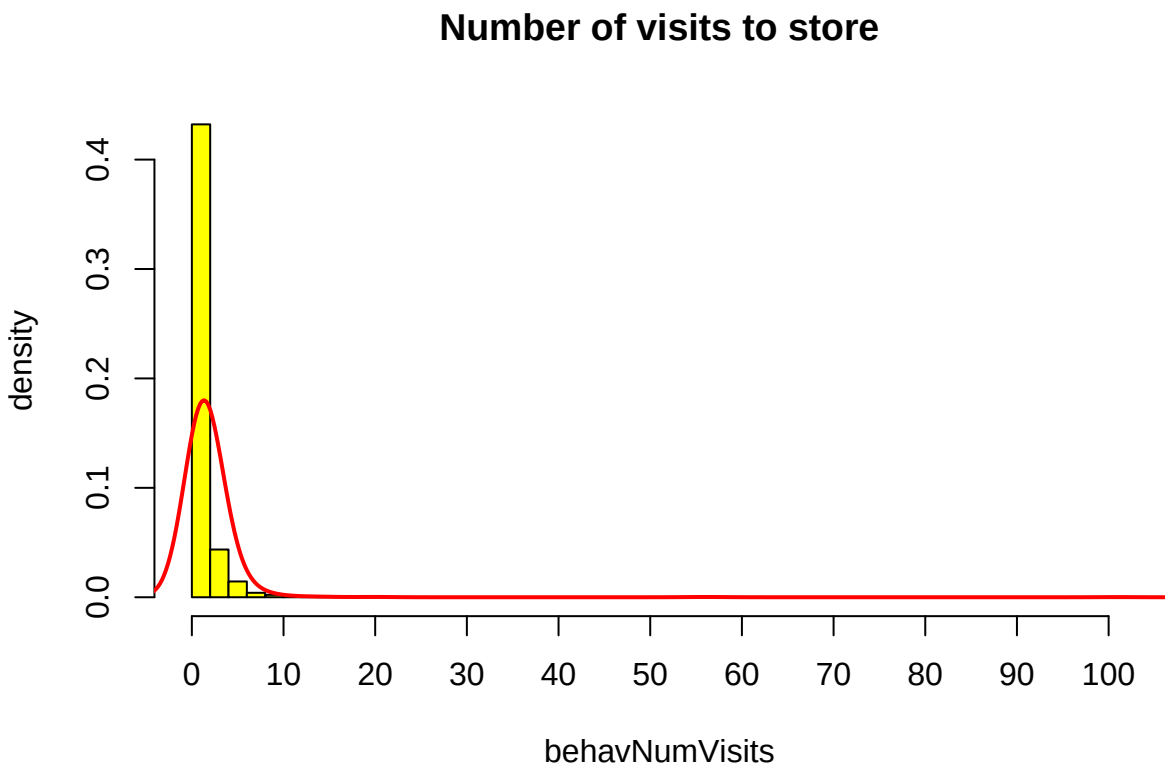
```
## TX
##  1
```

```r
# Question 7

# Getting the range of the variable behavNumVisits
range(ecomm.df$behavNumVisits)
```

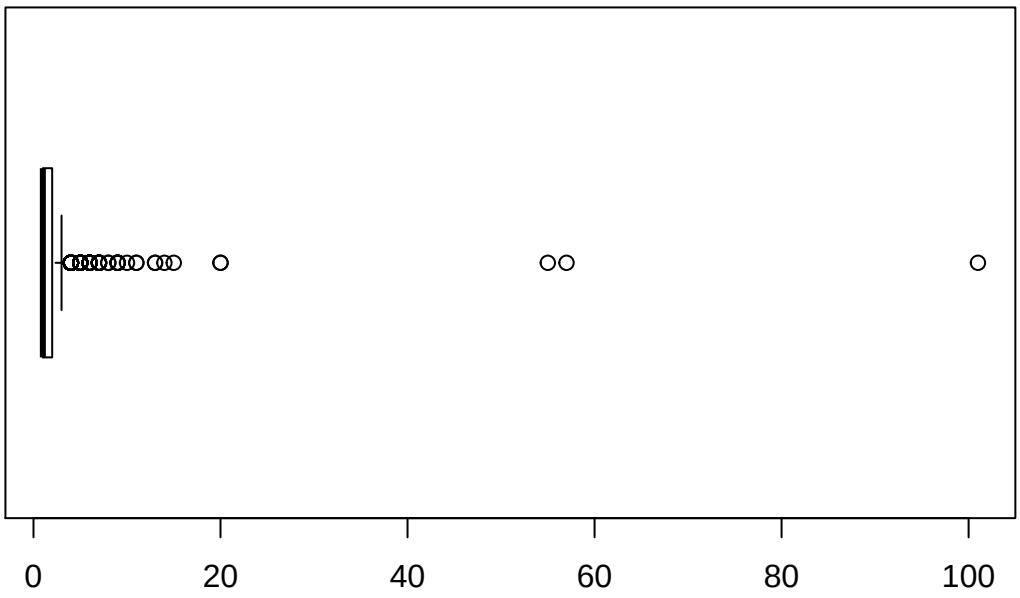
```
## [1]   1 101
```

```r
# Creating the histogram to show density with axis lablel, title and color edited and x axis broken int
hist(ecomm.df$behavNumVisits,xlab = 'behavNumVisits',main="Number of visits to store",ylab='density',co
# Custom setting of the x-axis ticks
axis(side=1,at=seq(0,100,by=10))
# Plotting a density line with adjusted smoothing and line width
lines(density(ecomm.df$behavNumVisits,bw=2),type="l", col="red", lwd=2)
```



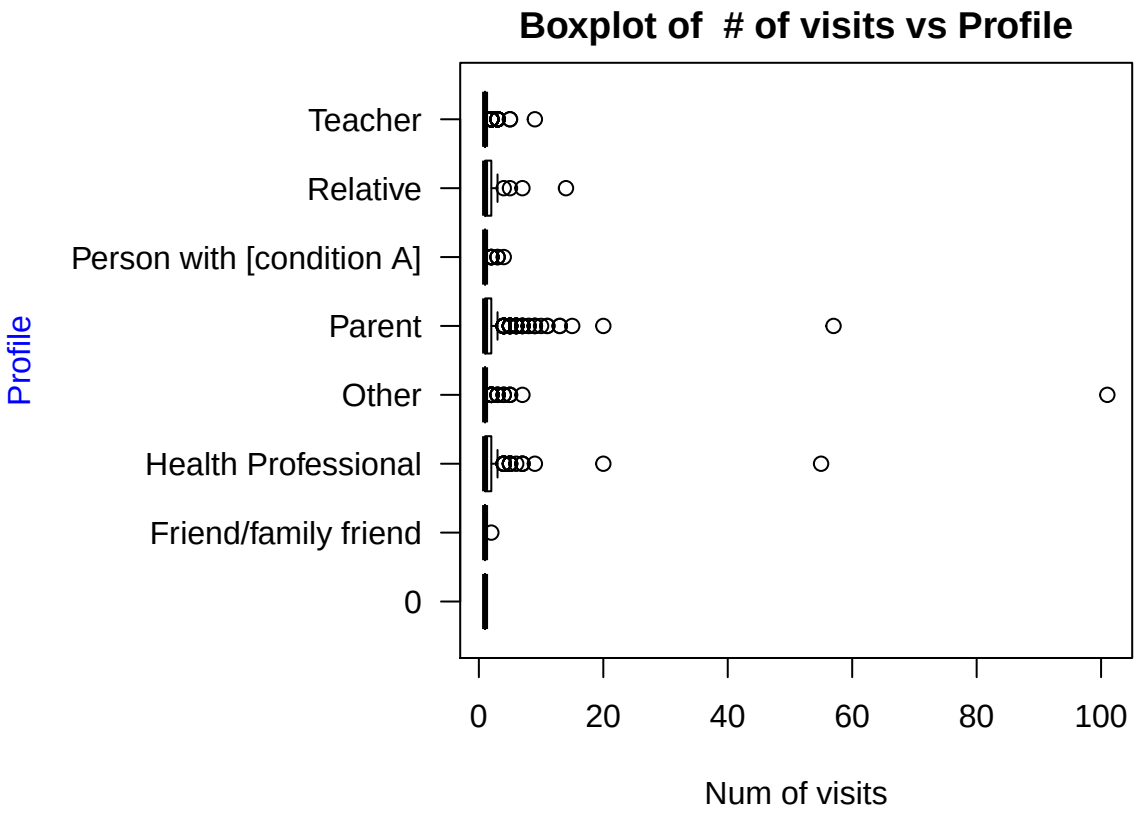

```
# Question 8

# Plotting a horizontal box plot for behavNumVisits
boxplot(ecomm.df$behavNumVisits,horizontal = TRUE)
```




```
# Question 9

#Setting the margin to show the chart with Ylabel and profile overlap
par(mar=c(5,13, 2, 2))
# Plotting a box plot for behavNumVisits by Profile
boxplot(ecomm.df$behavNumVisits~ecomm.df$profile,horizontal = TRUE,xlab = "Num of visits",ylab = "",mai
mtext("Profile",side=2,line=11,col="blue1")
```




```
# Resetting the margin to default
par(mar=c(5, 4, 4, 2)+ 0.1)
```