Homework 1

1. The Iowa data set iowa.csv is a toy example that summarises the yield of wheat (bushels per acre) for the state of Iowa between 1930-1962. In addition to yield, year, rainfall and temperature were recorded as the main predictors of yield.
   a. First, we need to load the data set into R using the command `read.csv()`. Use the help function to learn what arguments this function takes. Once you have the necessary input, load the data set into R and make it a data frame called `iowa.df`.
   b. How many rows and columns does `iowa.df` have?
   c. What are the names of the columns of `iowa.df`?
   d. What is the value of row 5, column 7 of `iowa.df`?
   e. Display the second row of `iowa.df` in its entirety.

```
setwd("D:/github/Rlab/")
Iowa.df<-read.csv("data/Iowa.csv",header=T)
Iowa.df = as.data.frame(Iowa.df)
names(Iowa.df)
```

```
##  [1] "Year"  "Rain0" "Temp1" "Rain1" "Temp2" "Rain2" "Temp3" "Rain3" "Temp4"
## [10] "Yield"
```

```
Iowa.df[5,7]
```

```
## [1] 79.7
```

```
Iowa.df[2,]
```

```
##   Year Rain0 Temp1 Rain1 Temp2 Rain2 Temp3 Rain3 Temp4 Yield
## 2 1930  14.8  57.5  3.83    75  2.72  77.2   3.3  72.6  32.9
```

2. Syntax and class-typing.

a. For each of the following commands, either explain why they should be errors, or explain the non-erroneous result.

vector1 <- c(“5”, “12”, “7”, “32”):non-erroneous because this is a simple assignment statement, c() Combine Values into a Vector or List, vector1 is a vector with 4 charactors.

max(vector1):non-erroneous because charactor can compare to each other, according to ascii code.

sort(vector1):non-erroneous because charactor can compare to each other, according to ascii code.

sum(vector1): error because charactor can not be add. sum function only take numeric or complex or logical vectors as input.

b. For the next series of commands, either explain their results, or why they should produce errors.

vector2 <- c(“5”,7,12):output a vector with 3 charactor whose value are “5” “7” “12”. the out put type of c() is determined from the highest type of the components in the hierarchy NULL < raw < logical < integer < double < complex < character < list < expression, and all arguments are coerced to a common type which is the type of the returned value.

vector2[2] + vector2[3]:error, Arithmetic Operators perform arithmetic on numeric or complex vectors.

dataframe3 <- data.frame(z1=“5”,z2=7,z3=12):output a dataframe shape of 1*3,the first is a integer, the last two are double type. because data.frame() will change charactor type to integer type.

dataframe3[1,2] + dataframe3[1,3]:19, adding two double value.

list4 <- list(z1="6", z2=42, z3="49", z4=126):list with 4 values, z1 and z3 are charactors, z2 and z4 are double. list() will not change the type of data.

list4[[2]]+list4[[4]]:168, list4[[i]] means the value of the ith value.

list4[2]+list4[4]:error, list4[i] means the name and value of ith element, whose type is list.

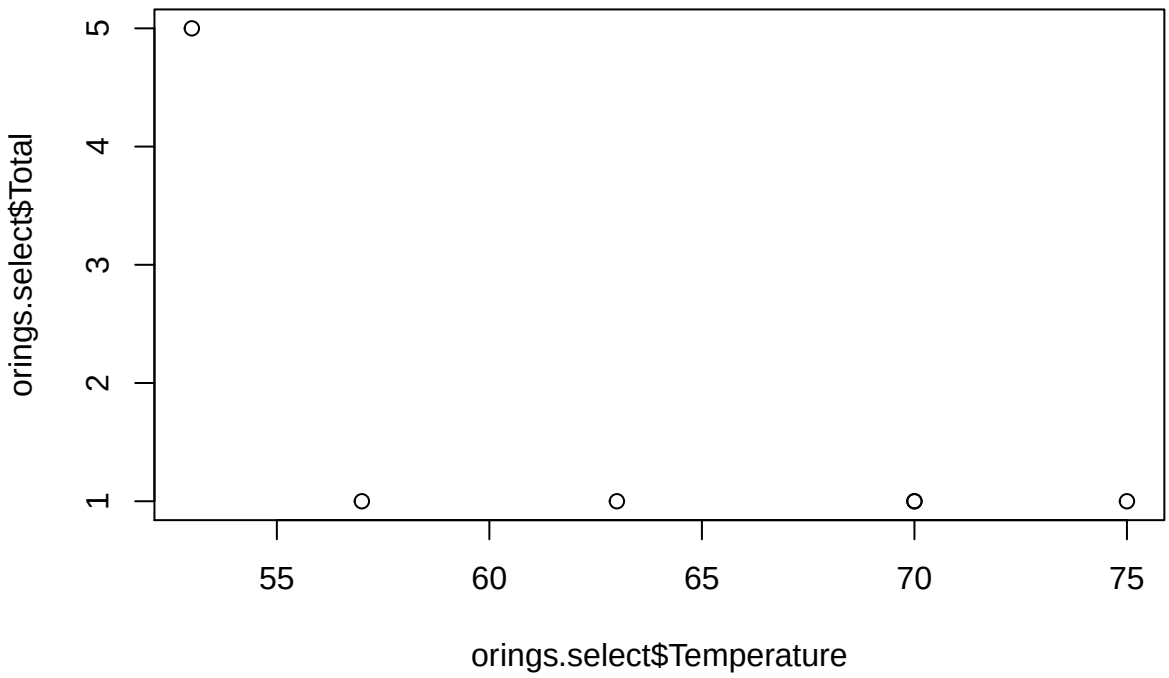
3. Working with functions and operators.
    a. The colon operator will create a sequence of integers in order. It is a special case of the function `seq()` which you saw earlier in this assignment. Using the help command `?seq` to learn about the function, design an expression that will give you the sequence of numbers from 1 to 10000 in increments of 372. Design another that will give you a sequence between 1 and 10000 that is exactly 50 numbers in length.
    b. The function `rep()` repeats a vector some number of times. Explain the difference between 'rep(1:3, times=3) and rep(1:3, each=3).

```
seq1 = seq(from=1, to = 10000, by = 372)
seq2 = seq(from = 1, to=10000, length.out = 50)
```

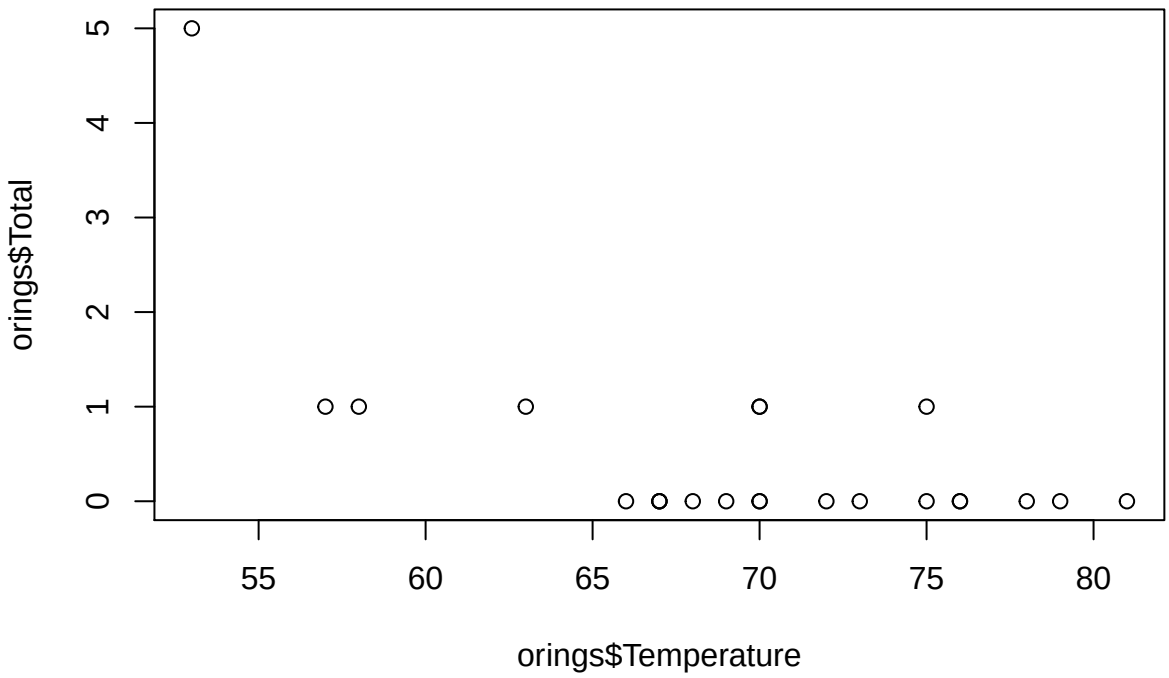
MB.Ch1.2. The orings data frame gives data on the damage that had occurred in US space shuttle launches prior to the disastrous Challenger launch of 28 January 1986. The observations in rows 1, 2, 4, 11, 13, and 18 were included in the pre-launch charts used in deciding whether to proceed with the launch, while remaining rows were omitted.

Create a new data frame by extracting these rows from orings, and plot total incidents against temperature for this new data frame. Obtain a similar plot for the full data set.

```
library(DAAG)
orings = as.data.frame(orings)
orings.select = orings[c(1,2,4,11,13,18),]
plot(orings.select$Temperature, orings.select$Total)
```




```
plot(orings$Temperature, orings$Total)
```



MB.Ch1.4. For the data frame ais (DAAG package)

(a) Use the function str() to get information on each of the columns. Determine whether any of the columns hold missing values.

```
library(DAAG)
str(ais)
```

```
## 'data.frame':    202 obs. of  13 variables:
##  $ rcc   : num  3.96 4.41 4.14 4.11 4.45 4.1 4.31 4.42 4.3 4.51 ...
##  $ wcc   : num  7.5 8.3 5 5.3 6.8 4.4 5.3 5.7 8.9 4.4 ...
##  $ hc    : num  37.5 38.2 36.4 37.3 41.5 37.4 39.6 39.9 41.1 41.6 ...
##  $ hg    : num  12.3 12.7 11.6 12.6 14 12.5 12.8 13.2 13.5 12.7 ...
##  $ ferr  : num  60 68 21 69 29 42 73 44 41 44 ...
##  $ bmi   : num  20.6 20.7 21.9 21.9 19 ...
##  $ ssf   : num  109.1 102.8 104.6 126.4 80.3 ...
##  $ pcBfat: num  19.8 21.3 19.9 23.7 17.6 ...
##  $ lbm   : num  63.3 58.5 55.4 57.2 53.2 ...
##  $ ht    : num  196 190 178 185 185 ...
##  $ wt    : num  78.9 74.4 69.1 74.9 64.6 63.7 75.2 62.3 66.5 62.9 ...
##  $ sex   : Factor w/ 2 levels "f","m": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sport : Factor w/ 10 levels "B_Ball","Field",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
names(ais)[which(colSums(apply(ais,c(1,2),is.na))>0)]#no column hold missing values
```

```
## character(0)
```

(b) Make a table that shows the numbers of males and females for each different sport. In which sports is there a large imbalance (e.g., by a factor of more than 2:1) in the numbers of the two sexes?

```
table = matrix(c(1,2,3),nrow = 1)
colnames(table) <- c("sports","m","f")
table = as.data.frame(table)
for(i in 1:dim(ais)[1]){
    if(any(ais$sport[i]==table$sports)){
            table[which(table$sports==ais$sport[i]),as.character(ais$sex[i])] =
                table[which(table$sports==ais$sport[i]),as.character(ais$sex[i])] + 1
    }
    else{
        table = rbind(table, list(sports=ais$sport[i],m=0,f=0))
        table[which(table$sports==ais$sport[i]),as.character(ais$sex[i])] =
            table[which(table$sports==ais$sport[i]),as.character(ais$sex[i])] + 1
    }
}
table = table[2:11,]
for(i in 1:10){
    if(table[i,2]>2*table[i,3]|table[i,3]>2*table[i,2]){
        print(table[i,1])
    }
}
```

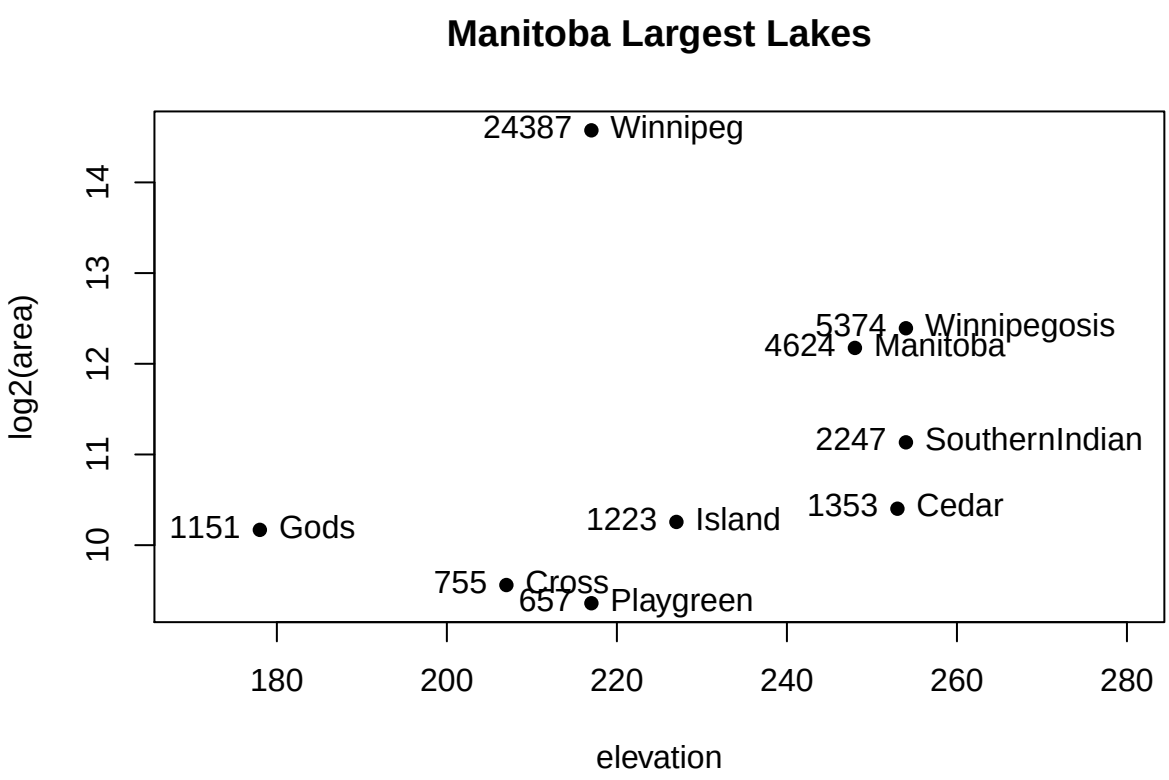
```
## [1] "Netball"
## [1] "T_Sprnt"
## [1] "Gym"
## [1] "W_Polo"
```

MB.Ch1.6.Create a data frame called Manitoba.lakes that contains the lake's elevation (in meters above sea level) and area (in square kilometers) as listed below. Assign the names of the lakes using the row.names() function. elevation area Winnipeg 217 24387 Winnipegosis 254 5374 Manitoba 248 4624 SouthernIndian 254 2247 Cedar 253 1353 Island 227 1223 Gods 178 1151 Cross 207 755 Playgreen 217 657

```
Manitoba.lakes = data.frame(Manitoba.lakes)
```

(a) Use the following code to plot log2(area) versus elevation, adding labeling infor- mation (there is an extreme value of area that makes a logarithmic scale pretty much essential):

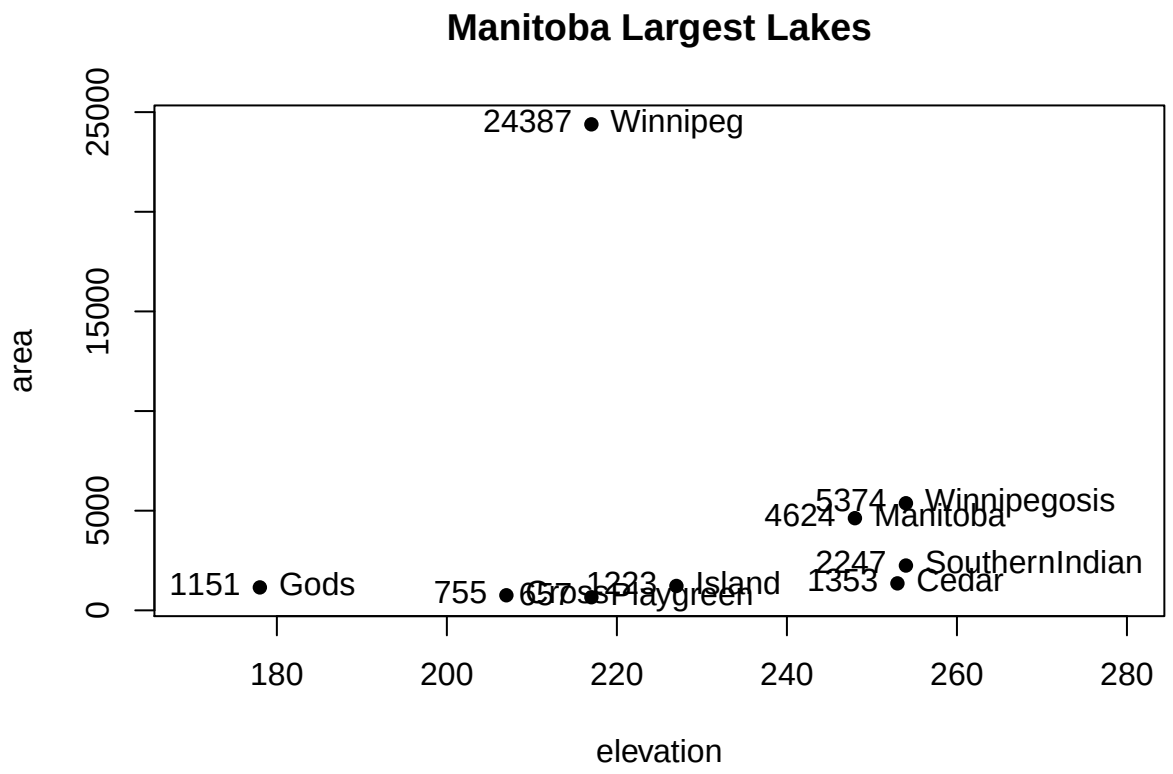
```
area = Manitoba.lakes$area
elevation = Manitoba.lakes$elevation
plot(log2(area) ~ elevation, pch=16, xlim=c(170,280))
# NB: Doubling the area increases log2(area) by 1.0
text(log2(area) ~ elevation, labels=row.names(Manitoba.lakes), pos=4)
text(log2(area) ~ elevation, labels=area, pos=2)
title("Manitoba Largest Lakes")
```



The labeling on the y-axis is the Base two logarithm of labeling on the points. Double the area, then the scale on the plot will add 1.

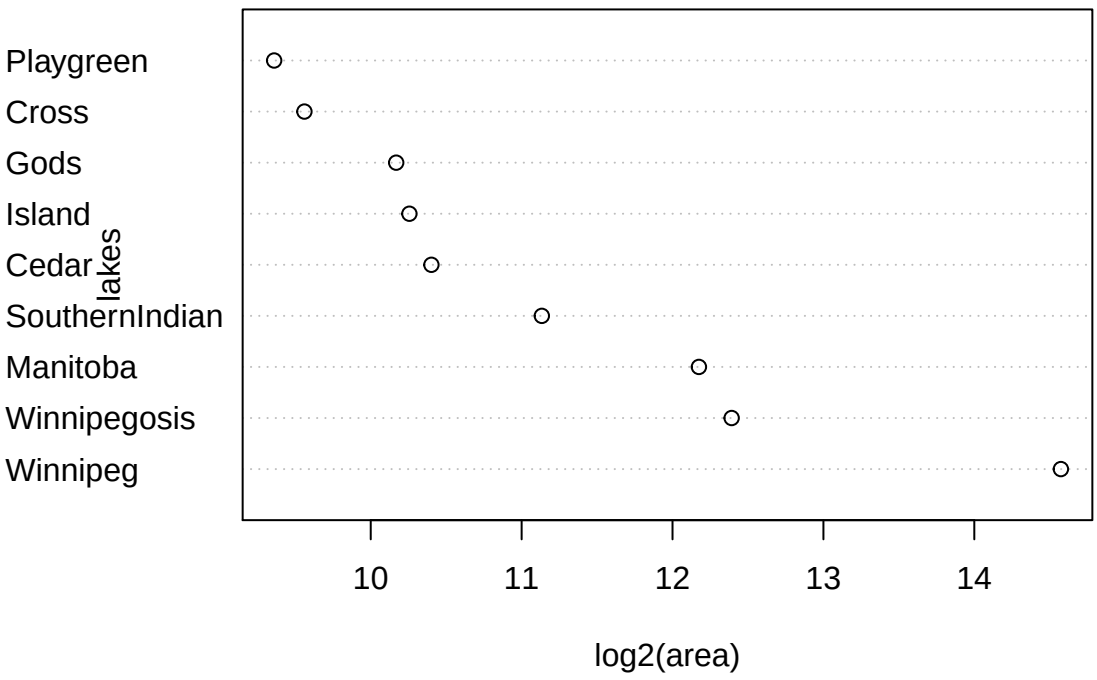
(b) Repeat the plot and associated labeling, now plotting area versus elevation, but specifying log="y" in order to obtain a logarithmic y-scale.

```
plot(area ~ elevation, pch=16, xlim=c(170,280), ylog=T)
text(area ~ elevation, labels=row.names(Manitoba.lakes), pos=4, ylog=T)
text(area ~ elevation, labels=area, pos=2, ylog=T)
title("Manitoba Largest Lakes")
```



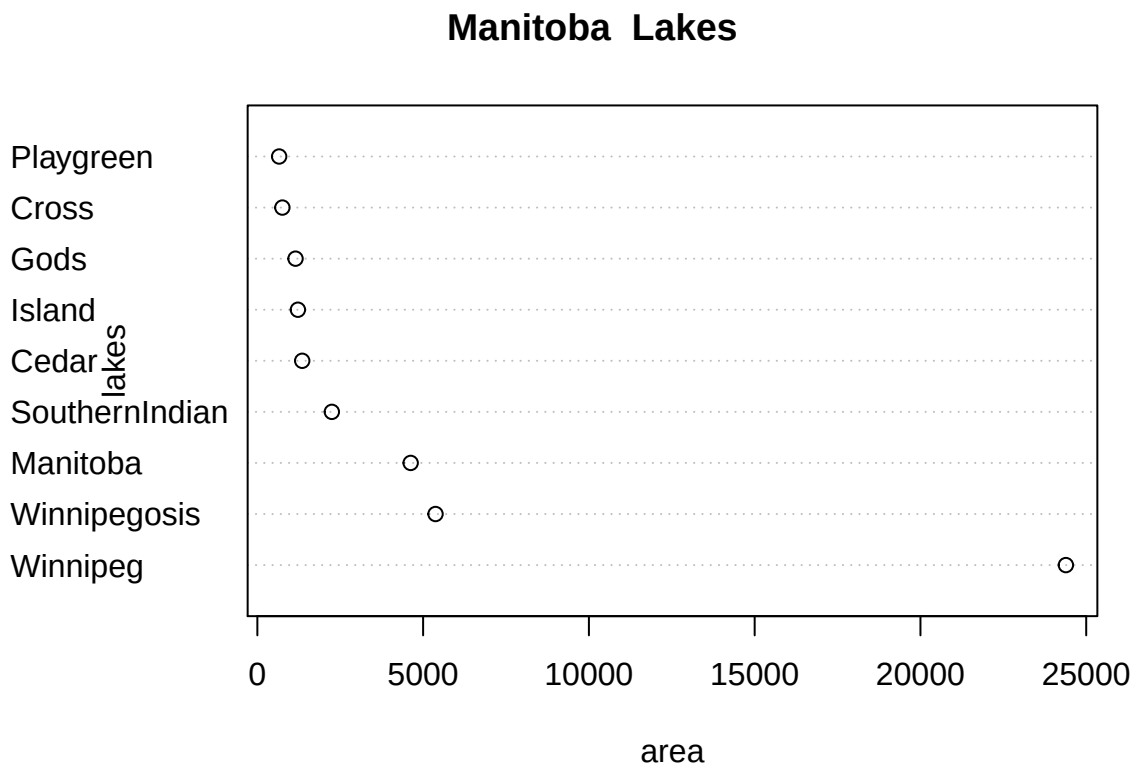
MB.Ch1.7. Look up the help page for the R function dotchart(). Use this function to display the areas of the Manitoba lakes (a) on a linear scale, and (b) on a logarithmic scale. Add, in each case, suitable labeling information.

```
dotchart(log2(Manitoba.lakes$area),labels=row.names(Manitoba.lakes),xlab = 'log2(area)', ylab = 'lakes')
title("Manitoba Lakes")
```




```
dotchart(Manitoba.lakes$area,labels=row.names(Manitoba.lakes),xlab = 'area', ylab = 'lakes')

title("Manitoba  Lakes")
```



MB.Ch1.8. Using the sum() function, obtain a lower bound for the area of Manitoba covered by water.

```
sum(Manitoba.lakes$area)
```

```
## [1] 41771
```