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.
2. 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.

**解**:

iowa.df<-read.csv("data/iowa.csv",header=T,sep=';')

1. How many rows and columns does iowa.df have?

**解**:

dim(iowa.df)

## [1] 33 10

1. What are the names of the columns of iowa.df?

**解**:

names(iowa.df)

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

1. What is the value of row 5, column 7 of iowa.df?

**解**:

iowa.df[5,7]

## [1] 79.7

1. Display the second row of iowa.df in its entirety.

**解**:

iowa.df[2,]

## Year Rain0 Temp1 Rain1 Temp2 Rain2 Temp3 Rain3 Temp4 Yield  
## 2 1931 14.76 57.5 3.83 75 2.72 77.2 3.3 72.6 32.9

1. Syntax and class-typing.
2. 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")  
max(vector1)  
sort(vector1)  
sum(vector1)

**解**：由于vector1中的数据格式为字符型，所以只能进行对字符类型的操作。max(vector1)的结果为“7”。sort(vector1)的结果为“12” “32” “5” “7”。而sum()只能对数字型数据才有效，所以将会报错。

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

vector2 <- c("5",7,12)  
vector2[2] + vector2[3]  
  
dataframe3 <- data.frame(z1="5",z2=7,z3=12)  
dataframe3[1,2] + dataframe3[1,3]  
  
list4 <- list(z1="6", z2=42, z3="49", z4=126)  
list4[[2]]+list4[[4]]  
list4[2]+list4[4]

**解**:在用c()赋值时，如果元素中含有字符型数据，则将所有非字符数据都转化为字符型，所以vector2=“5” “7” “12”，而对于字符串是没有加法操作的。

在用data.frame()进行数据框赋值时，数据类型不变，所以[1,2]和[1,3]的数据类型都是数字类型，可以进行dataframe3[1,2]+dataframe3[1,3]，结果为19。

在list数据中，用list[i]调用的是其第i个组分，仍然是一个列表，所以无法做加法，而list[[i]]访问的是第i个组分的元素值，如这里list4[[2]]就是元素42，所以可以进行加法，list4[[2]]+list[[4]]=168.

1. Working with functions and operators.
2. 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.

**解**：

seq(from=1,to=10000,by=372)

## [1] 1 373 745 1117 1489 1861 2233 2605 2977 3349 3721 4093 4465 4837 5209  
## [16] 5581 5953 6325 6697 7069 7441 7813 8185 8557 8929 9301 9673

seq(from=1,to=10000,length.out=50)

## [1] 1.0000 205.0612 409.1224 613.1837 817.2449 1021.3061  
## [7] 1225.3673 1429.4286 1633.4898 1837.5510 2041.6122 2245.6735  
## [13] 2449.7347 2653.7959 2857.8571 3061.9184 3265.9796 3470.0408  
## [19] 3674.1020 3878.1633 4082.2245 4286.2857 4490.3469 4694.4082  
## [25] 4898.4694 5102.5306 5306.5918 5510.6531 5714.7143 5918.7755  
## [31] 6122.8367 6326.8980 6530.9592 6735.0204 6939.0816 7143.1429  
## [37] 7347.2041 7551.2653 7755.3265 7959.3878 8163.4490 8367.5102  
## [43] 8571.5714 8775.6327 8979.6939 9183.7551 9387.8163 9591.8776  
## [49] 9795.9388 10000.0000

1. 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).

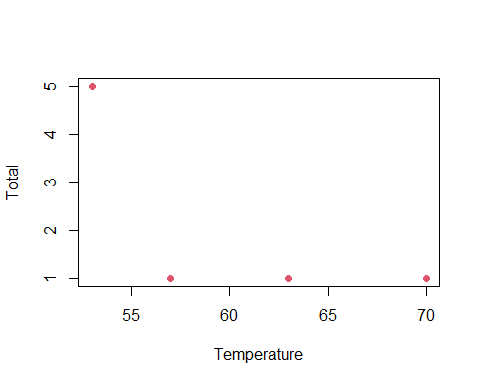
**解**:前者是将序列整体重复3遍，结果是123123123；后者是将每个元素分别重复三遍，结果为111222333。

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.

**解**：

newdata1=orings[c(1,2,4,11,13),]  
plot(Total~Temperature,data=newdata1,col=2,pch=16,cex=1)



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

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

**解**:

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 ...

for(i in (1:dim(ais)[2])){  
 vec<-is.na(ais[,i])  
 if(any(vec)){  
 cat("the column",i,"hold missing values")  
 }  
}

结果没有输出显示，所以没有列含有NA缺省值。

1. 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?

**解**：

table(Sex=ais$sex,Sports=ais$sport)

## Sports  
## Sex B\_Ball Field Gym Netball Row Swim T\_400m T\_Sprnt Tennis W\_Polo  
## f 13 7 4 23 22 9 11 4 7 0  
## m 12 12 0 0 15 13 18 11 4 17

其中Gym,Netball,T\_Sprnt,W\_polo男女相差比较大。

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.

**解**：

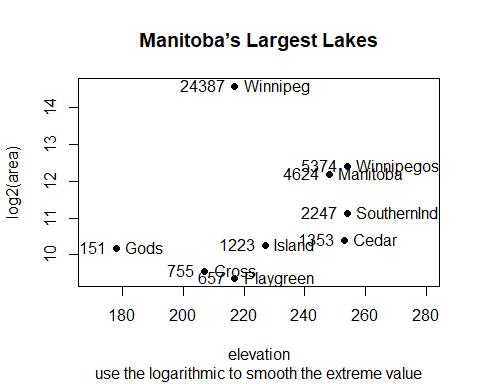
ele<-c(217,254,248,254,253,227,178,207,217)  
are<-c(24387,5374,4624,2247,1353,1223,1151,755,657)  
Manitoba.lakes<-data.frame(elevation=ele,area=are)  
row.names(Manitoba.lakes)<-c("Winnipeg","Winnipegosis","Manitoba","SouthernIndian","Cedar","Island","Gods","Cross","Playgreen")  
Manitoba.lakes

## 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

1. 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):

**解**：

attach(Manitoba.lakes)  
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(main="Manitoba’s Largest Lakes",sub="use the logarithmic to smooth the extreme value")

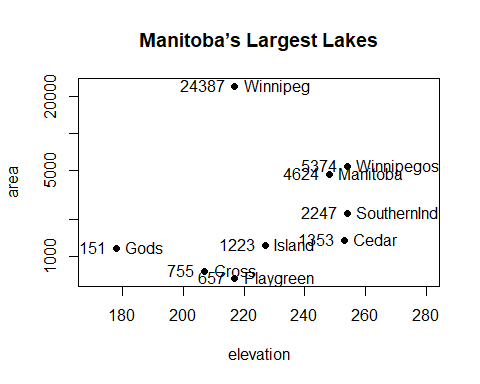


Devise captions that explain the labeling on the points and on the y-axis. It will be necessary to explain how distances on the scale relate to changes in area.

1. Repeat the plot and associated labeling, now plotting area versus elevation, but specifying log=“y” in order to obtain a logarithmic y-scale.

**解**：

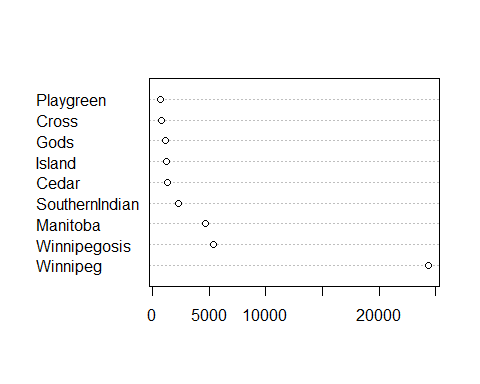
plot.default(area ~ elevation, pch=16, xlim=c(170,280),log="y")  
text.default(area ~ elevation, labels=row.names(Manitoba.lakes), pos=4)  
text.default(area ~ elevation, labels=area, pos=2)   
title("Manitoba’s Largest Lakes")



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.

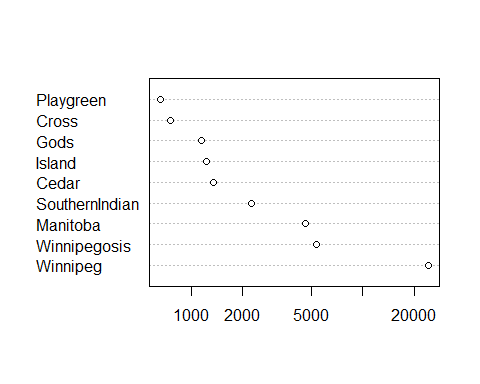
**解**： (a)

dotchart(area,row.names(Manitoba.lakes))



(b)

dotchart(area,row.names(Manitoba.lakes),log="x")



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

**解**：

sum(area)

## [1] 41771