Homework2

Yining Song

Problem3

Version control can show the details of improvements or differences between versions. Therefore, it can record our progress in the classroom in a step by step manner, which will help us get a better understanding of the materials we need to learn, as well as inspire us of what needs improvements.

Problem 4

(a)

```
data1=read.csv("Sensory.dat.txt")
summary(data1)
```

```
##                    Operator
##  0.9 3.1 1.1 1.9 1.6   : 1
##  1 4.3 4.9 3.3 5.3 4.4 : 1
##  1.3 2.4 0.8 1.2 1.3   : 1
##  1.9 3.9 2.6 4.6 2.2   : 1
##  10 5.0 4.8 3.9 5.5 3.8: 1
##  2 6.0 5.3 4.5 5.9 4.7 : 1
##  (Other)               :25
```

Obviously the data were messed up, and the title did not match the data. So we skip the title line:

```
data1=read.table("Sensory.dat.txt",fill = T,skip = 2)
print(data1)
```

```
##      V1  V2  V3  V4  V5  V6
## 1   1.0 4.3 4.9 3.3 5.3 4.4
## 2   4.3 4.5 4.0 5.5 3.3  NA
## 3   4.1 5.3 3.4 5.7 4.7  NA
## 4   2.0 6.0 5.3 4.5 5.9 4.7
## 5   4.9 6.3 4.2 5.5 4.9  NA
## 6   6.0 5.9 4.7 6.3 4.6  NA
## 7   3.0 2.4 2.5 2.3 3.1 2.4
## 8   3.9 3.0 2.8 2.7 1.3  NA
## 9   1.9 3.9 2.6 4.6 2.2  NA
## 10  4.0 7.4 8.2 6.4 6.8 6.0
## 11  7.1 7.9 5.9 7.3 6.1  NA
## 12  6.4 7.1 6.9 7.0 6.7  NA
## 13  5.0 5.7 6.3 5.4 6.1 5.9
## 14  5.8 5.7 5.4 6.2 6.5  NA
## 15  5.8 6.0 6.1 7.0 4.9  NA
## 16  6.0 2.2 2.4 1.7 3.4 1.7
## 17  3.0 1.8 2.1 4.0 1.7  NA
## 18  2.1 3.3 1.1 3.3 2.1  NA
## 19  7.0 1.2 1.5 1.2 0.9 0.7
## 20  1.3 2.4 0.8 1.2 1.3  NA
## 21  0.9 3.1 1.1 1.9 1.6  NA
## 22  8.0 4.2 4.8 4.5 4.6 3.2
## 23  3.0 4.5 4.7 4.9 4.6  NA
## 24  4.8 4.8 4.7 4.8 4.3  NA
```

```
## 25  9.0 8.0 8.6 9.0 9.4 8.8
## 26  9.0 7.7 6.7 9.0 7.9  NA
## 27  8.9 9.2 8.1 9.1 7.6  NA
## 28 10.0 5.0 4.8 3.9 5.5 3.8
## 29  5.4 5.0 3.4 4.9 4.6  NA
## 30  2.8 5.2 4.1 3.9 5.5  NA
```

We do not need the indicators 1 to 10, so we remove them from data:

```
data1[1,]=data1[1,][-1]
data1[4,]=data1[4,][-1]
data1[7,]=data1[7,][-1]
data1[10,]=data1[10,][-1]
data1[13,]=data1[13,][-1]
data1[16,]=data1[16,][-1]
data1[19,]=data1[19,][-1]
data1[22,]=data1[22,][-1]
data1[25,]=data1[25,][-1]
data1[28,]=data1[28,][-1]
data1=data1[,-6]
names(data1)=c('1','2','3','4','5')
tidydata1=data1
print(tidydata1)
```

```
##      1   2   3   4   5
## 1  4.3 4.9 3.3 5.3 4.4
## 2  4.3 4.5 4.0 5.5 3.3
## 3  4.1 5.3 3.4 5.7 4.7
## 4  6.0 5.3 4.5 5.9 4.7
## 5  4.9 6.3 4.2 5.5 4.9
## 6  6.0 5.9 4.7 6.3 4.6
## 7  2.4 2.5 2.3 3.1 2.4
## 8  3.9 3.0 2.8 2.7 1.3
## 9  1.9 3.9 2.6 4.6 2.2
## 10 7.4 8.2 6.4 6.8 6.0
## 11 7.1 7.9 5.9 7.3 6.1
## 12 6.4 7.1 6.9 7.0 6.7
## 13 5.7 6.3 5.4 6.1 5.9
## 14 5.8 5.7 5.4 6.2 6.5
## 15 5.8 6.0 6.1 7.0 4.9
## 16 2.2 2.4 1.7 3.4 1.7
## 17 3.0 1.8 2.1 4.0 1.7
## 18 2.1 3.3 1.1 3.3 2.1
## 19 1.2 1.5 1.2 0.9 0.7
## 20 1.3 2.4 0.8 1.2 1.3
## 21 0.9 3.1 1.1 1.9 1.6
## 22 4.2 4.8 4.5 4.6 3.2
## 23 3.0 4.5 4.7 4.9 4.6
## 24 4.8 4.8 4.7 4.8 4.3
## 25 8.0 8.6 9.0 9.4 8.8
## 26 9.0 7.7 6.7 9.0 7.9
## 27 8.9 9.2 8.1 9.1 7.6
## 28 5.0 4.8 3.9 5.5 3.8
## 29 5.4 5.0 3.4 4.9 4.6
## 30 2.8 5.2 4.1 3.9 5.5
```

This is the tidy dataset.

```
summary(tidydata1)
```

```
##        1                2                3                4
##  Min.   :0.900   Min.   :1.500   Min.   :0.800   Min.   :0.900
##  1st Qu.:2.850   1st Qu.:3.450   1st Qu.:2.650   1st Qu.:3.925
##  Median :4.550   Median :4.950   Median :4.150   Median :5.400
##  Mean   :4.593   Mean   :5.063   Mean   :4.167   Mean   :5.193
##  3rd Qu.:5.950   3rd Qu.:6.225   3rd Qu.:5.400   3rd Qu.:6.275
##  Max.   :9.000   Max.   :9.200   Max.   :9.000   Max.   :9.400
##        5
##  Min.   :0.700
##  1st Qu.:2.250
##  Median :4.600
##  Mean   :4.267
##  3rd Qu.:5.800
##  Max.   :8.800
```

(b)

```
data2=read.csv("LongJumpData.dat.txt")
summary(data2)
```

```
##           Year.Long.Jump.Year.Long.Jump.Year.Long.Jump.Year.Long.Jump
##  -4 249.75 24 293.13 56 308.25 80 336.25:1
##  0 282.88 28 304.75 60 319.75 84 336.25 :1
##  12 299.25 48 308.00 72 324.50          :1
##  20 281.50 52 298.00 76 328.50          :1
##  4 289.00 32 300.75 64 317.75 88 343.25 :1
##  8 294.50 36 317.31 68 350.50 92 342.50 :1
```

Obviously there are missing values in the table, and the title did not match the data. So we fill out the missing values and skip the title line:

```
data2=read.table("LongJumpData.dat.txt",fill = T,skip = 1)
print(data2)
```

```
##   V1     V2 V3     V4 V5     V6 V7     V8
## 1 -4 249.75 24 293.13 56 308.25 80 336.25
## 2  0 282.88 28 304.75 60 319.75 84 336.25
## 3  4 289.00 32 300.75 64 317.75 88 343.25
## 4  8 294.50 36 317.31 68 350.50 92 342.50
## 5 12 299.25 48 308.00 72 324.50 NA     NA
## 6 20 281.50 52 298.00 76 328.50 NA     NA
```

Now we summary the years and the performances in two columns:

```
Year=c(data2[,1],data2[,3],data2[,5],data2[,7])
Performance=c(data2[,2],data2[,4],data2[,6],data2[,8])
tidydata2=cbind(Year,Performance)[-c(23,24),]
tidydata2[,1]=tidydata2[,1]+1900
print(tidydata2)
```

```
##       Year Performance
##  [1,] 1896      249.75
##  [2,] 1900      282.88
##  [3,] 1904      289.00
```

```
##  [4,] 1908      294.50
##  [5,] 1912      299.25
##  [6,] 1920      281.50
##  [7,] 1924      293.13
##  [8,] 1928      304.75
##  [9,] 1932      300.75
## [10,] 1936      317.31
## [11,] 1948      308.00
## [12,] 1952      298.00
## [13,] 1956      308.25
## [14,] 1960      319.75
## [15,] 1964      317.75
## [16,] 1968      350.50
## [17,] 1972      324.50
## [18,] 1976      328.50
## [19,] 1980      336.25
## [20,] 1984      336.25
## [21,] 1988      343.25
## [22,] 1992      342.50
```

This is the tidy dataset.

```
summary(tidydata2)
```

```
##       Year       Performance
##  Min.   :1896   Min.   :249.8
##  1st Qu.:1921   1st Qu.:295.4
##  Median :1950   Median :308.1
##  Mean   :1945   Mean   :310.3
##  3rd Qu.:1971   3rd Qu.:327.5
##  Max.   :1992   Max.   :350.5
```

(c)

```
data3=read.csv("BrainandBodyWeight.dat.txt")
print(data3)
```

```
##    Body.Wt.Brain.Wt.Body.Wt.Brain.Wt
## 1                3.385 44.5 521.000 655.0 2.500 12.10
## 2                  0.480 15.5 0.785 3.5 55.500 175.00
## 3               1.350 8.1 10.000 115.0 100.000 157.00
## 4              465.000 423.0 3.300 25.6 52.160 440.00
## 5                36.330 119.5 0.200 5.0 10.550 179.50
## 6                  27.660 115.0 1.410 17.5 0.550 2.40
## 7              14.830 98.2 529.000 680.0 60.000 81.00
## 8                 1.040 5.5 207.000 406.0 3.600 21.00
## 9                 4.190 58.0 85.000 325.0 4.288 39.20
## 10                    0.425 6.4 0.750 12.3 0.280 1.90
## 11                0.101 4.0 62.000 1320.0 0.075 1.20
## 12               0.920 5.7 6654.000 5712.0 0.122 3.00
## 13                     1.000 6.6 3.500 3.9 0.048 0.33
## 14               0.005 0.1 6.800 179.0 192.000 180.00
## 15                  0.060 1.0 35.000 56.0 3.000 25.00
## 16               3.500 10.8 4.050 17.0 160.000 169.00
## 17                    2.000 12.3 0.120 1.0 0.900 2.60
## 18                    1.700 6.3 0.023 0.4 1.620 11.40
```

```
## 19               2547.000 4603.0 0.010 0.3 0.104 2.50
## 20                   0.023 0.3 1.400 12.5 4.235 50.40
## 21                       187.100 419.0 250.000 490.0
```

Obviously there are missing values in the table, and the title did not match the data. So we fill out the missing values and skip the title line:

```
data3=read.table("BrainandBodyWeight.dat.txt",fill = T,skip = 1)
print(data3)
```

```
##          V1     V2       V3     V4      V5     V6
## 1     3.385   44.5  521.000  655.0   2.500  12.10
## 2     0.480   15.5    0.785    3.5  55.500 175.00
## 3     1.350    8.1   10.000  115.0 100.000 157.00
## 4   465.000  423.0    3.300   25.6  52.160 440.00
## 5    36.330  119.5    0.200    5.0  10.550 179.50
## 6    27.660  115.0    1.410   17.5   0.550   2.40
## 7    14.830   98.2  529.000  680.0  60.000  81.00
## 8     1.040    5.5  207.000  406.0   3.600  21.00
## 9     4.190   58.0   85.000  325.0   4.288  39.20
## 10    0.425    6.4    0.750   12.3   0.280   1.90
## 11    0.101    4.0   62.000 1320.0   0.075   1.20
## 12    0.920    5.7 6654.000 5712.0   0.122   3.00
## 13    1.000    6.6    3.500    3.9   0.048   0.33
## 14    0.005    0.1    6.800  179.0 192.000 180.00
## 15    0.060    1.0   35.000   56.0   3.000  25.00
## 16    3.500   10.8    4.050   17.0 160.000 169.00
## 17    2.000   12.3    0.120    1.0   0.900   2.60
## 18    1.700    6.3    0.023    0.4   1.620  11.40
## 19 2547.000 4603.0    0.010    0.3   0.104   2.50
## 20    0.023    0.3    1.400   12.5   4.235  50.40
## 21  187.100  419.0  250.000  490.0      NA     NA
```

Now we summary the body weights and the brain weights in two columns:

```
BodyW=c(data3[,1],data3[,3],data3[,5])
BrainW=c(data3[,2],data3[,4],data3[,6])
tidydata3=cbind(BrainW,BodyW)[-63,]
print(tidydata3)
```

```
##        BrainW    BodyW
##  [1,]   44.50    3.385
##  [2,]   15.50    0.480
##  [3,]    8.10    1.350
##  [4,]  423.00  465.000
##  [5,]  119.50   36.330
##  [6,]  115.00   27.660
##  [7,]   98.20   14.830
##  [8,]    5.50    1.040
##  [9,]   58.00    4.190
## [10,]    6.40    0.425
## [11,]    4.00    0.101
## [12,]    5.70    0.920
## [13,]    6.60    1.000
## [14,]    0.10    0.005
## [15,]    1.00    0.060
```

```
## [16,]   10.80    3.500
## [17,]   12.30    2.000
## [18,]    6.30    1.700
## [19,] 4603.00 2547.000
## [20,]    0.30    0.023
## [21,]  419.00  187.100
## [22,]  655.00  521.000
## [23,]    3.50    0.785
## [24,]  115.00   10.000
## [25,]   25.60    3.300
## [26,]    5.00    0.200
## [27,]   17.50    1.410
## [28,]  680.00  529.000
## [29,]  406.00  207.000
## [30,]  325.00   85.000
## [31,]   12.30    0.750
## [32,] 1320.00   62.000
## [33,] 5712.00 6654.000
## [34,]    3.90    3.500
## [35,]  179.00    6.800
## [36,]   56.00   35.000
## [37,]   17.00    4.050
## [38,]    1.00    0.120
## [39,]    0.40    0.023
## [40,]    0.30    0.010
## [41,]   12.50    1.400
## [42,]  490.00  250.000
## [43,]   12.10    2.500
## [44,]  175.00   55.500
## [45,]  157.00  100.000
## [46,]  440.00   52.160
## [47,]  179.50   10.550
## [48,]    2.40    0.550
## [49,]   81.00   60.000
## [50,]   21.00    3.600
## [51,]   39.20    4.288
## [52,]    1.90    0.280
## [53,]    1.20    0.075
## [54,]    3.00    0.122
## [55,]    0.33    0.048
## [56,]  180.00  192.000
## [57,]   25.00    3.000
## [58,]  169.00  160.000
## [59,]    2.60    0.900
## [60,]   11.40    1.620
## [61,]    2.50    0.104
## [62,]   50.40    4.235
```

This is the tidy dataset.

```
summary(tidydata3)
```

```
##      BrainW            BodyW
##  Min.   :   0.10   Min.   :   0.005
##  1st Qu.:   4.25   1st Qu.:   0.600
```

```
##  Median :  17.25   Median :   3.342
##  Mean   : 283.13   Mean   : 198.790
##  3rd Qu.: 166.00   3rd Qu.:  48.203
##  Max.   :5712.00   Max.   :6654.000
```

(d) By viewing the data in the txt file, we can see that we have data in 6 treatment levels, each have 3 data points. So we first store them in 6 vectors manually.

```
Ife1=c(16.1,15.3,17.5)
Ife2=c(16.6,19.2,18.5)
Ife3=c(20.8,18.0,21.0)
PED1=c(8.1,8.6,10.1)
PED2=c(12.7,13.7,11.5)
PED3=c(14.4,15.4,13.7)
Yield=c(Ife1,Ife2,Ife3,PED1,PED2,PED3)
treatments=c(rep(1,3),rep(2,3),rep(3,3),rep(4,3),rep(5,3),rep(6,3))
tidydata4=cbind(Yield,treatments)
print(tidydata4)
```

```
##       Yield treatments
##  [1,]  16.1          1
##  [2,]  15.3          1
##  [3,]  17.5          1
##  [4,]  16.6          2
##  [5,]  19.2          2
##  [6,]  18.5          2
##  [7,]  20.8          3
##  [8,]  18.0          3
##  [9,]  21.0          3
## [10,]   8.1          4
## [11,]   8.6          4
## [12,]  10.1          4
## [13,]  12.7          5
## [14,]  13.7          5
## [15,]  11.5          5
## [16,]  14.4          6
## [17,]  15.4          6
## [18,]  13.7          6
```

Problem 5

```
library(swirl)
```

```
##
## | Hi! I see that you have some variables saved in your workspace. To keep
## | things running smoothly, I recommend you clean up before starting swirl.
##
## | Type ls() to see a list of the variables in your workspace. Then, type
## | rm(list=ls()) to clear your workspace.
##
## | Type swirl() when you are ready to begin.
```

```
.datapath <- file.path(path.package('swirl'), 'Courses',
                       'R_Programming_E', 'Looking_at_Data',
                       'plant-data.txt')
plants <- read.csv(.datapath, strip.white=TRUE, na.strings="")
.cols2rm <- c('Accepted.Symbol', 'Synonym.Symbol')
```

```
plants <- plants[, !(names(plants) %in% .cols2rm)]
names(plants) <- c('Scientific_Name', 'Duration', 'Active_Growth_Period',
                   'Foliage_Color', 'pH_Min', 'pH_Max',
                   'Precip_Min', 'Precip_Max',
                   'Shade_Tolerance', 'Temp_Min_F')
```