HW2-Min

Jie Min

9/10/2019

Problem 3

One way version control can help me is that wherever I am and whenever I want, I can review my previous codes. I can easily download my previous code and continue doing my work when I am outside home or school.

Problem 4

a

First several rows of the cleaned data and the summarize of the cleaned data is showed below:

```
url<-"http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/Sensory.dat"
data1<-read.table(url,header=FALSE,skip=2,fill=TRUE)
data1<-cbind(data1,matrix(0,ncol=1,nrow=nrow(data1)))
idx<-seq(1,30,3)
data1[idx,2:7]<-data1[idx,1:6]
data1[-idx,3:7]<-data1[-idx,1:5]
data1<-data1[,-2]
data1[,1]<-sort(rep(1:10,3))
colnames(data1)<-c('Item','1','2','3','4','5')
```

```
##   Item   1   2   3   4   5
## 1    1 4.3 4.9 3.3 5.3 4.4
## 2    1 4.3 4.5 4.0 5.5 3.3
## 3    1 4.1 5.3 3.4 5.7 4.7
## 4    2 6.0 5.3 4.5 5.9 4.7
## 5    2 4.9 6.3 4.2 5.5 4.9
## 6    2 6.0 5.9 4.7 6.3 4.6
```

```
##        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
```

The data is the observed value of 10 itmes on 5 operators, however in the raw data, not each row has the number of items, so some lines have 5 numbers and some lines have 6 numbers in the origin dataset.

I first filled the blank as NA when importing the data to make sure each row has same number of columns.Then I right align the dataframe, and filled the missed item number into the first column for lines that don't have this value.

The issue of the uncleaned data is that it has missing values in 'Item' column. I didn't see any issue in the cleaned data.

b

First several rows of the cleaned data and the summarize of the cleaned data is showed below:

```
url<-"http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/LongJumpData.dat"
data1<-read.table(url,skip=1,fill=TRUE)
data1<-as.matrix(cbind(data1,matrix(NA,ncol=1,nrow=nrow(data1))))
data2<-rbind(data1[,c(2,3)],data1[,c(4,5)],data1[,c(6,7)],data1[,c(8,9)])
data2<-as.data.frame(cbind(rep(data1[,1],4),data2))
colnames(data2)=c('Year','Long','Jump')
data2<-data2[order(data2$Year),]
data2[,1]<-data2[,1]+1990
```

```
##   Year   Long Jump
## 1 1986 249.75   24
## 2 1986 293.13   56
## 3 1986 308.25   80
## 4 1986 336.25   NA
## 5 1990 282.88   28
## 6 1990 304.75   60
```

```
##       Long            Jump
##  Min.   :249.8   Min.   :24
##  1st Qu.:295.4   1st Qu.:45
##  Median :308.1   Median :62
##  Mean   :310.3   Mean   :60
##  3rd Qu.:327.5   3rd Qu.:77
##  Max.   :350.5   Max.   :92
##  NA's   :2       NA's   :8
```

For this data, I gathered columns that are data from the same year into rows. Finally I have three columns in the cleaned data table, the first column is year, the second column is Long and the third column is Jump.

The issue of this data is that it has several Na values. I didn't remove those Na's, but when calculating summary statistics, I omitted those Na's, and only used data we have.

c

```
url<-'http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/BrainandBodyWeight.dat'
data1<-as.matrix(unname(read.table(url,skip=1,fill=TRUE)))
data1<-cbind(seq(1,21,1),data1)
data2<-rbind(data1[,c(2,3)],data1[,c(4,5)],data1[,c(6,7)])
data2<-as.data.frame(cbind(rep(seq(1,21,1),3),data2))
colnames(data2)<-c('N','BodyWt','BrainWt')
data2<-data2[order(data2$N),]
data2<-data2[-63,-1]
```

```
##    BodyWt BrainWt
## 1   3.385    44.5
## 2 521.000   655.0
## 3   2.500    12.1
## 4   0.480    15.5
## 5   0.785     3.5
## 6  55.500   175.0
```

```
##      BodyWt            BrainWt
##  Min.   :   0.005   Min.   :   0.10
##  1st Qu.:   0.600   1st Qu.:   4.25
##  Median :   3.342   Median :  17.25
##  Mean   : 198.790   Mean   : 283.13
##  3rd Qu.:  48.203   3rd Qu.: 166.00
##  Max.   :6654.000   Max.   :5712.00
```

For this data, I gathered columns that are data into rows.I have two columns in the cleaned data table, the first column is BodyWt, the second column is BrainWt.

There are several issues about this data. First, there is no value about which bodywt and brainwt are for which species. Second, the recorded value may have some problem. For example, for the value in the first row, bodywt is much more smaller than brainwt is not reasonable.

d

```
url<-'http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/tomato.dat'


l<- readLines(url)[c(3,4)]
l<-gsub(" ",',',l)
data1<-as.matrix(read.table(text=l,comment.char = '*',sep=',',fill=TRUE)[,-1])
id1<-which(is.na(data1[1,])==TRUE)
id2<-which(is.na(data1[2,])==TRUE)
data2<-as.data.frame(rbind(data1[1,-id1],data1[2,-id2]))
```

```
##                10000 10000 10000 20000 20000 20000 30000 30000 30000
## Ife#1           16.1  15.3  17.5  16.6  19.2  18.5  20.8  18.0  21.0
## PusaEarlyDwarf   8.1   8.6  10.1  12.7  13.7  11.5  14.4  15.4  13.7
```

```
##      10000           10000           10000           20000
##  Min.   : 8.1   Min.   : 8.60   Min.   :10.10   Min.   :12.70
##  1st Qu.:10.1   1st Qu.:10.28   1st Qu.:11.95   1st Qu.:13.68
##  Median :12.1   Median :11.95   Median :13.80   Median :14.65
##  Mean   :12.1   Mean   :11.95   Mean   :13.80   Mean   :14.65
##  3rd Qu.:14.1   3rd Qu.:13.62   3rd Qu.:15.65   3rd Qu.:15.62
##  Max.   :16.1   Max.   :15.30   Max.   :17.50   Max.   :16.60
##      20000           20000           30000           30000
##  Min.   :13.70   Min.   :11.50   Min.   :14.4   Min.   :15.40
##  1st Qu.:15.07   1st Qu.:13.25   1st Qu.:16.0   1st Qu.:16.05
##  Median :16.45   Median :15.00   Median :17.6   Median :16.70
##  Mean   :16.45   Mean   :15.00   Mean   :17.6   Mean   :16.70
##  3rd Qu.:17.82   3rd Qu.:16.75   3rd Qu.:19.2   3rd Qu.:17.35
##  Max.   :19.20   Max.   :18.50   Max.   :20.8   Max.   :18.00
##      30000
```

```
##  Min.   :13.70
##  1st Qu.:15.53
##  Median :17.35
##  Mean   :17.35
##  3rd Qu.:19.18
##  Max.   :21.00
```

When reading the table, I changed the separate character from "" to ",". Then after reading in the table, I removed all blank cells, and finally got a clean dataset with two rows and nine columns.

The issue of original data is that data are separated both by ',' and by blank, thus it is difficult to read into r. I didn't see any issue in the cleaned data.

Problem 5

First, omit all rows that have NA values. Then we use the average value as a statistic that combines the information of pH_Min and pH_Max. A summary of the data is listed below.

```
plants1<-na.omit(plants)
plants1$pH<-(plants1$pH_Max+plants1$pH_Min)/2
```

```
##       Scientific_Name  Duration Active_Growth_Period Foliage_Color pH_Min
## 4      Abies balsamea Perennial    Spring and Summer         Green    4.0
## 9   Acacia constricta Perennial    Spring and Summer         Green    7.0
## 14 Acalypha virginica    Annual Spring, Summer, Fall         Green    5.9
## 17       Acer negundo Perennial    Spring and Summer         Green    5.0
## 19        Acer nigrum Perennial    Spring and Summer         Green    4.5
## 20 Acer pensylvanicum Perennial    Spring and Summer         Green    4.4
##    pH_Max Precip_Min Precip_Max Shade_Tolerance Temp_Min_F   pH
## 4     6.0         13         60        Tolerant        -43 5.00
## 9     8.5          4         20      Intolerant        -13 7.75
## 14    7.0         13         60    Intermediate         33 6.45
## 17    7.8         15         75        Tolerant        -46 6.40
## 19    7.3         24         60        Tolerant        -47 5.90
## 20    6.5         24         76        Tolerant        -47 5.45
```

```
##            Scientific_Name                         Duration
##  Abies balsamea    :  1    Perennial                   :692
##  Acacia constricta :  1    Annual                      : 64
##  Acalypha virginica:  1    Annual, Perennial           : 33
##  Acer negundo      :  1    Annual, Biennial            :  8
##  Acer nigrum       :  1    Annual, Biennial, Perennial:  6
##  Acer pensylvanicum:  1    Biennial, Perennial         :  6
##  (Other)           :807    (Other)                     :  4
##               Active_Growth_Period      Foliage_Color     pH_Min
##  Spring and Summer      :443       Dark Green  : 82   Min.   :3.000
##  Spring                 :143       Gray-Green  : 24   1st Qu.:4.500
##  Spring, Summer, Fall   : 90       Green       :675   Median :5.000
##  Summer                 : 87       Red         :  3   Mean   :4.988
##  Summer and Fall        : 20       White-Gray  :  9   3rd Qu.:5.500
##  Fall, Winter and Spring: 15       Yellow-Green: 20   Max.   :7.000
##  (Other)                : 15
##      pH_Max         Precip_Min      Precip_Max         Shade_Tolerance
```

```
##  Min.   : 5.100   Min.   : 4.00   Min.   : 16.00   Intermediate:239
##  1st Qu.: 7.000   1st Qu.:17.00   1st Qu.: 55.00   Intolerant  :332
##  Median : 7.300   Median :29.00   Median : 60.00   Tolerant    :242
##  Mean   : 7.335   Mean   :25.66   Mean   : 58.64
##  3rd Qu.: 7.700   3rd Qu.:32.00   3rd Qu.: 60.00
##  Max.   :10.000   Max.   :60.00   Max.   :200.00
##
##    Temp_Min_F           pH
##  Min.   :-79.00   Min.   :4.300
##  1st Qu.:-38.00   1st Qu.:5.800
##  Median :-33.00   Median :6.150
##  Mean   :-22.57   Mean   :6.161
##  3rd Qu.:-18.00   3rd Qu.:6.500
##  Max.   : 52.00   Max.   :8.200
##
```

Then by lm, we tested the relationship between Foliage_Color and pH. The coefficients and ANOVA results are listed in tables below. We can see that Foliage_Color and pH have strong relationship.

```
model1<-lm(pH~Foliage_Color,data=plants1)
model2<-anova(lm(pH~Foliage_Color,data=plants1))
```

Table 1: Coefficients

	x
(Intercept)	5.9993902
Foliage_ColorGray-Green	0.3714431
Foliage_ColorGreen	0.1757209
Foliage_ColorRed	0.4006098
Foliage_ColorWhite-Gray	0.4450542
Foliage_ColorYellow-Green	-0.0618902

Table 2: ANOVA TABLE

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Foliage_Color	5	5.226242	1.0452484	3.612788	0.0030772
Residuals	807	233.480517	0.2893191	NA	NA