Data Analysis Using R: Exercise2

罗智超 (ROKIA.ORG)　　1814347@qq.com

1 Data Cleaning

- Using file "data/survey2014_student.xls"

```
survey<-read.csv(file="data/survey2014_student.csv",header=TRUE,
                 sep=";", fileEncoding  = "GB2312",
                 numerals = "no.loss",
                 colClasses =
                c("numeric","character",
                  "numeric","numeric","numeric","numeric"))

dim(s)

s<-na.omit(survey)
summary(s)

boxplot(s[,3:6])
for (i in 3:6){
hist(s[,i],main = paste("Histogram of variable",colnames(s)[i]))
}
s<-s[s$height<300,]
nrow(s[s$weight<60,c(2:4)])

#update wrong weight data

#method1
```

```
s[s$weight<60,]$weight<-s[s$weight<60,]$weight*2

s[s$weight<80 & trimws(s$sex,which="both") ==" 男" &
    s$height>170,]$weight<-s[s$weight<80 &
  trimws(s$sex,which="both") ==" 男" &
    s$height>170,]$weight*2


#method2
library(dplyr)
s<-s %>% filter(weight<80 &
                  height>170 & trimws(sex)==" 男") %>%
                  mutate(weight=weight*2)
#method3
erow1<-ifelse(s$weight<60,TRUE,FALSE)
erow<-ifelse(s$weight<80 & s$height>170 & trimws(s$sex)==" 男",TRUE,FALSE)
s[erow,"weight"]<-s[erow,4]*2
s[erow1,"weight"]<-s[erow1,4]*2
```

2 Deal with Batch files

- Download dataset csv.rar

```
#Method1

fileName <- dir("D:/tempdata/csv")
scode<-substr(fileName,1,6)
nfile<-length(fileName)

for(i in 1:nfile){
assign(paste("s",scode[i], sep=""),
       read.csv(fileName[i],header=TRUE))
}
```

```
# Method2

fileName <- dir("D:/tempdata/csv")
cls <- c("character","character","character",
         "numeric","numeric","numeric","numeric")
stocklist<- lapply(fileName,function(x)
  read.csv(x,header=TRUE,colClasses=cls,stringsAsFactors=F))
allstcok<- do.call(rbind,stocklist)
```

3 Smoothing Binning

- Question:

```
x<-c(13,15,16,16,19,20,20,21,22,25,25,25,30,33,33,35,35,35,36,40,45,46,52,70)
# 生成 y 变量，其值为 x 中依次每三个数字的平均值（注：不是移动平均，是每三个数字算一个均值）
```