DATAGURU 炼数成金

数据分析、展现与R语言 第12周

2013.04.20

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

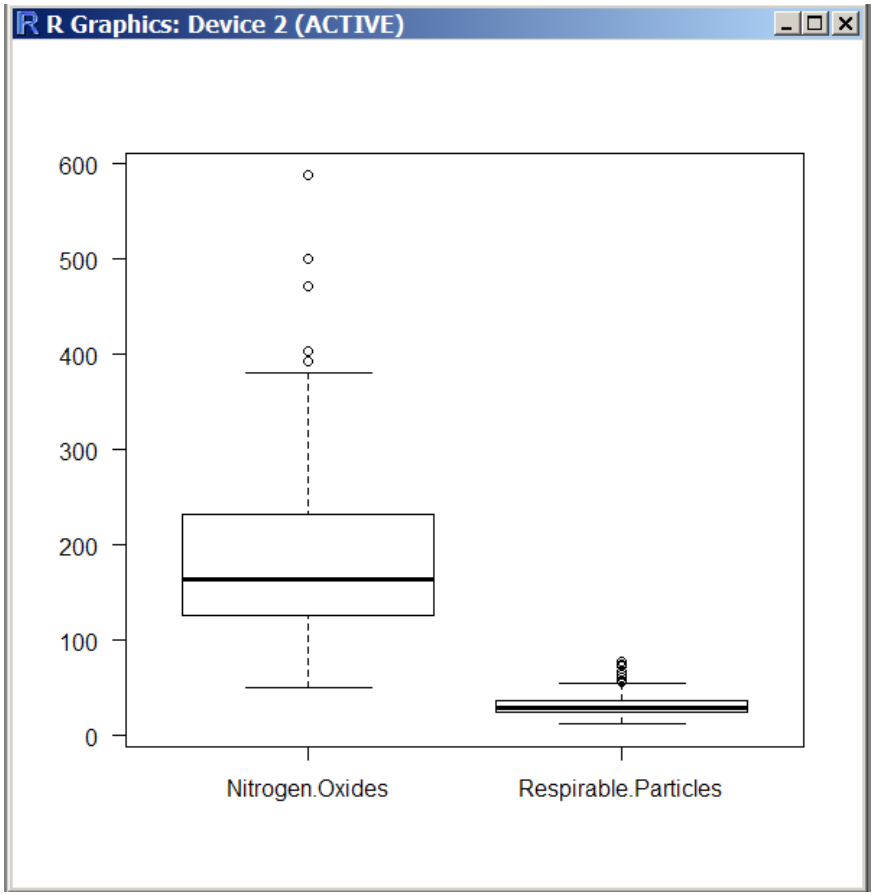
课程详情访问炼数成金培训网站

http://edu.dataguru.cn

箱型图

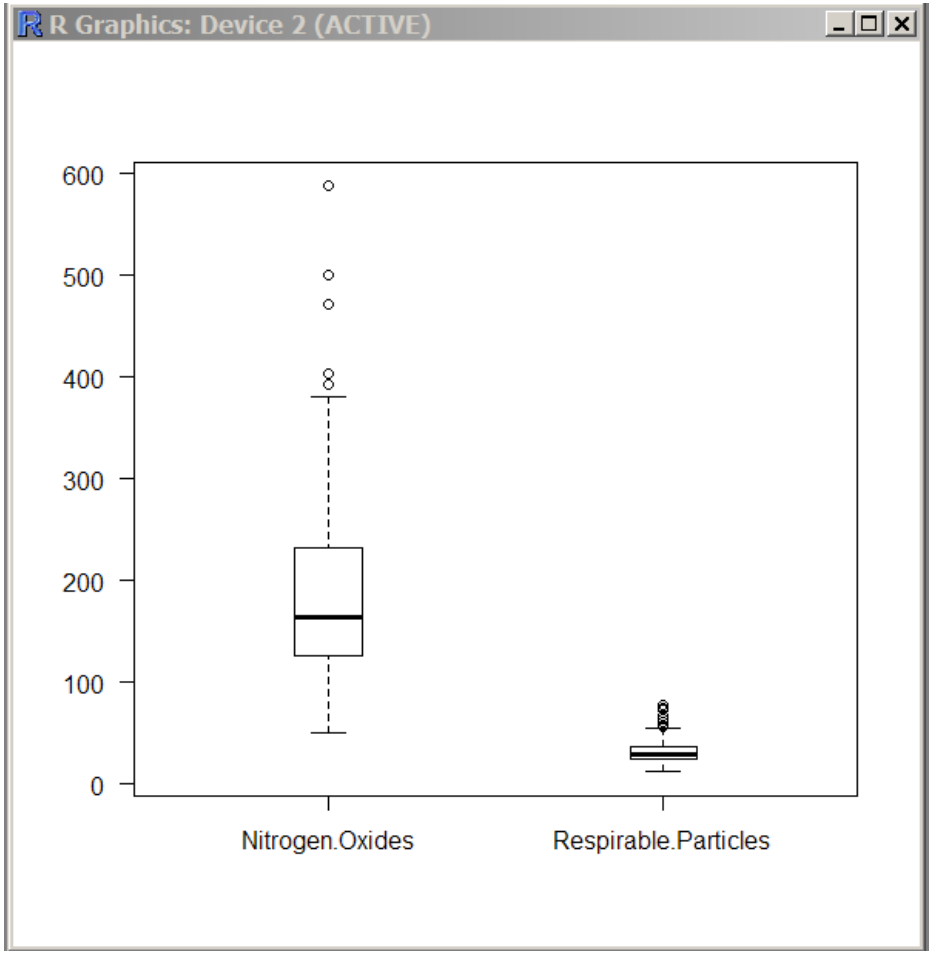
使用第七周数据

```
air<-read.csv("airpollution.csv")
boxplot(air,las=1)
```



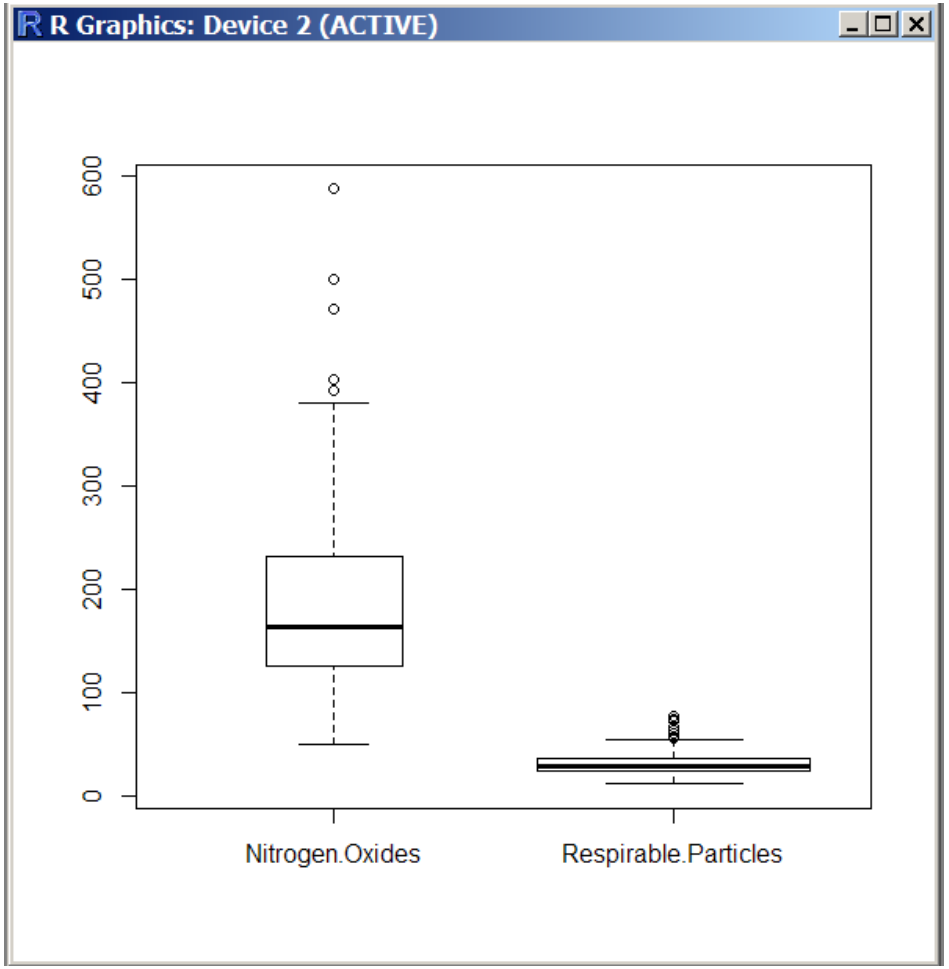
收窄箱体宽度

boxplot(air,boxwex=0.2,las=1)



指定箱体的宽度

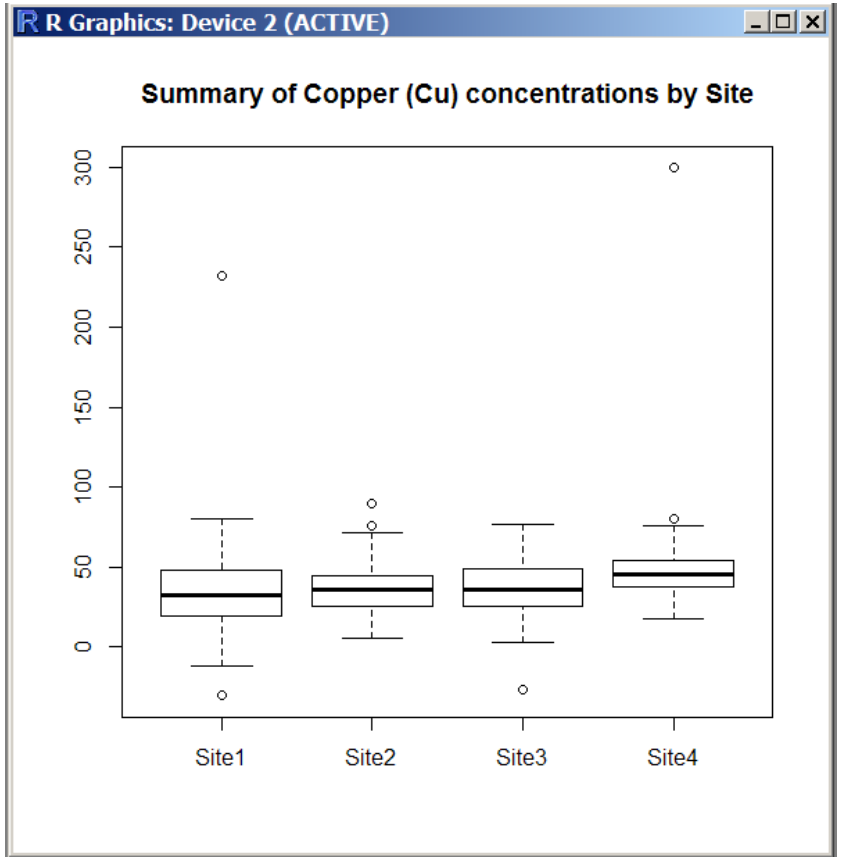
boxplot(air,width=c(1,2))



2013.04.20

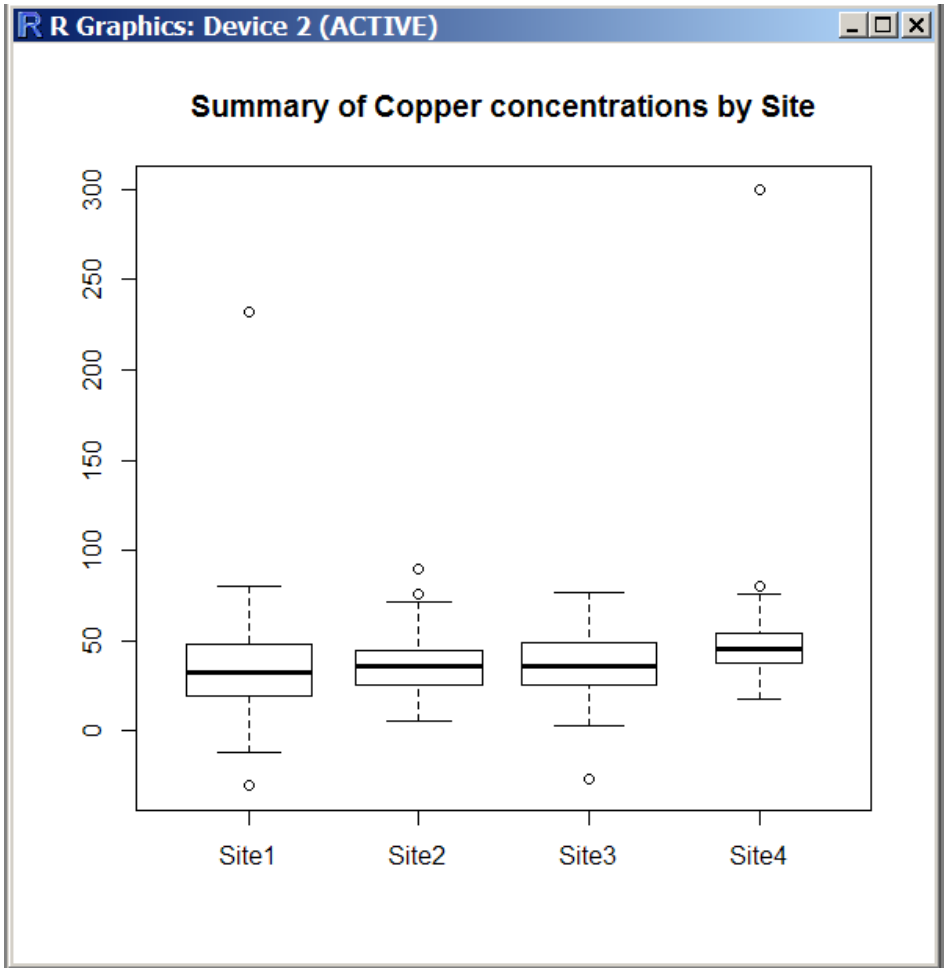
分组

```
metals<-read.csv("metals.csv")
boxplot(Cu~Source,data=metals,
main="Summary of Copper (Cu) concentrations by
      Site")
boxplot(Cu~Source*Expt,data=metals,
main="Summary of Copper (Cu) concentrations by
      Site")
```



观测值数量决定箱体宽度

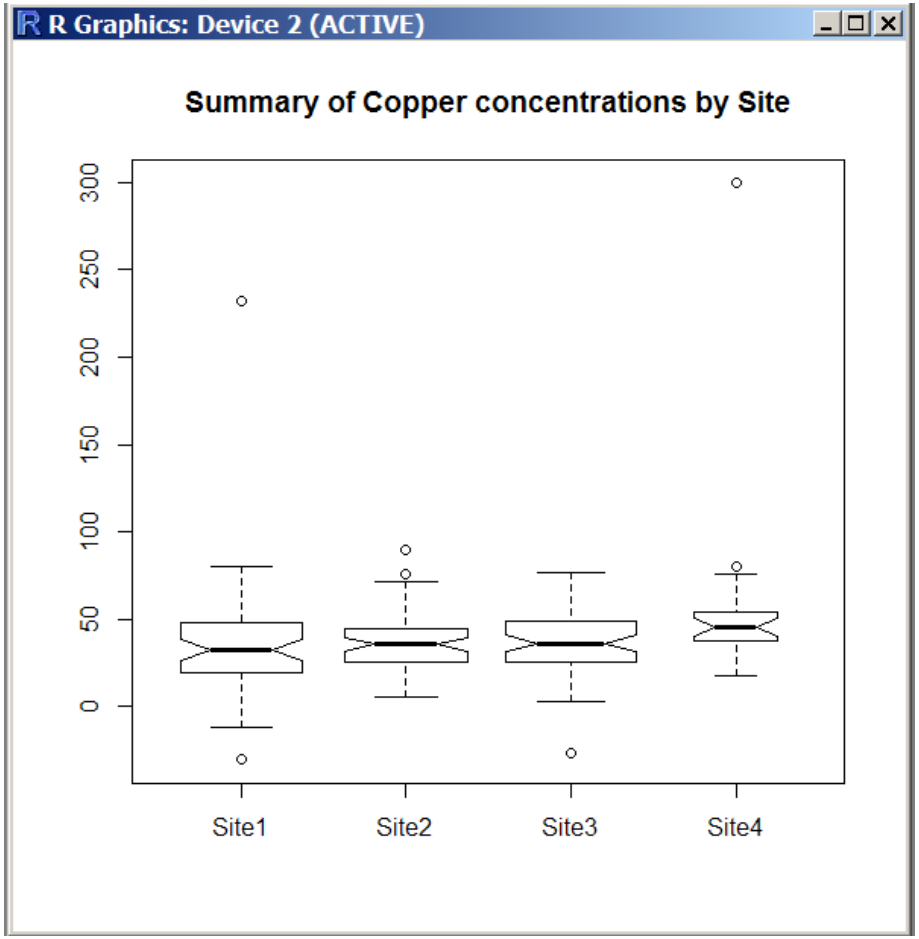
```
boxplot(Cu ~ Source, data =
    metals,varwidth=TRUE,
main="Summary of Copper
    concentrations by Site")
```



2013.04.20

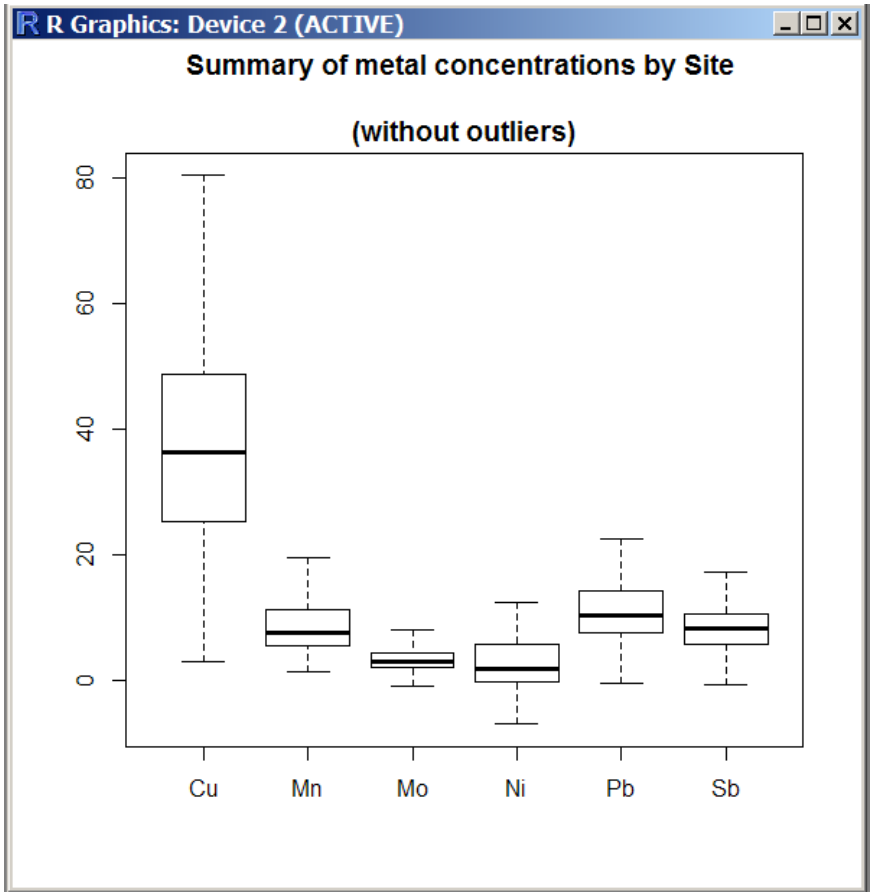
带notch的箱型图

```
boxplot(Cu ~ Source, data = metals,
varwidth=TRUE,notch=TRUE,
main="Summary of Copper
    concentrations by Site")
```



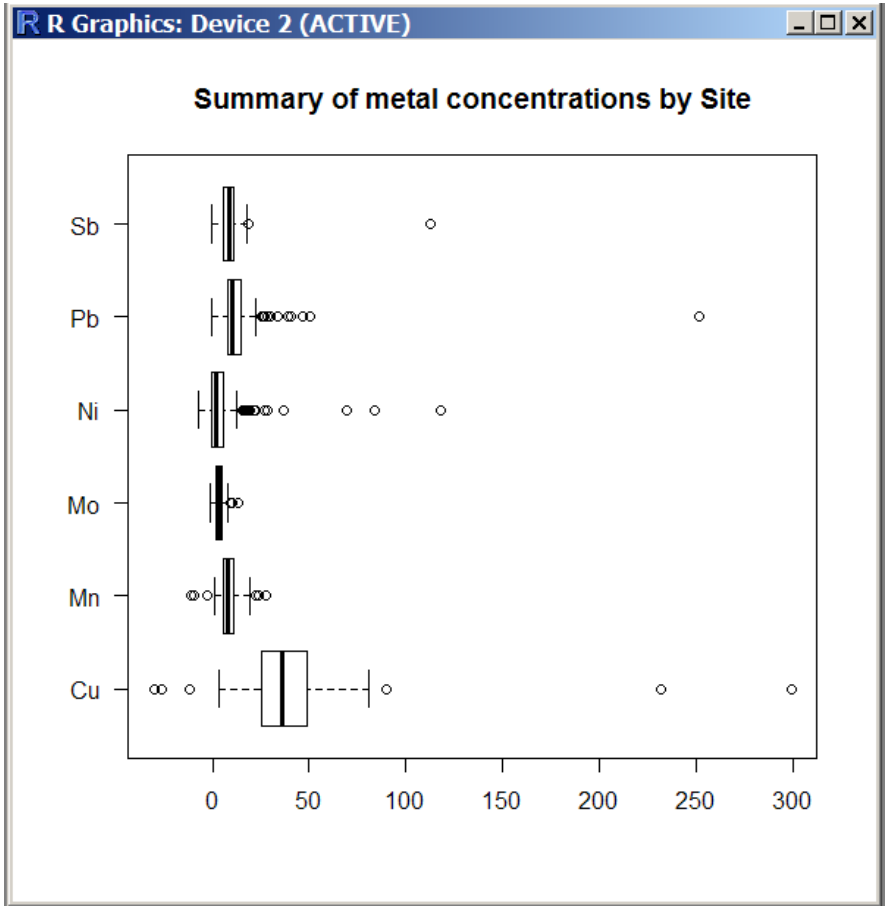
排除离群值

```
boxplot(metals[,-1],outline=FALSE,
main="Summary of metal
    concentrations by Site \n
(without outliers)")
```



水平放置

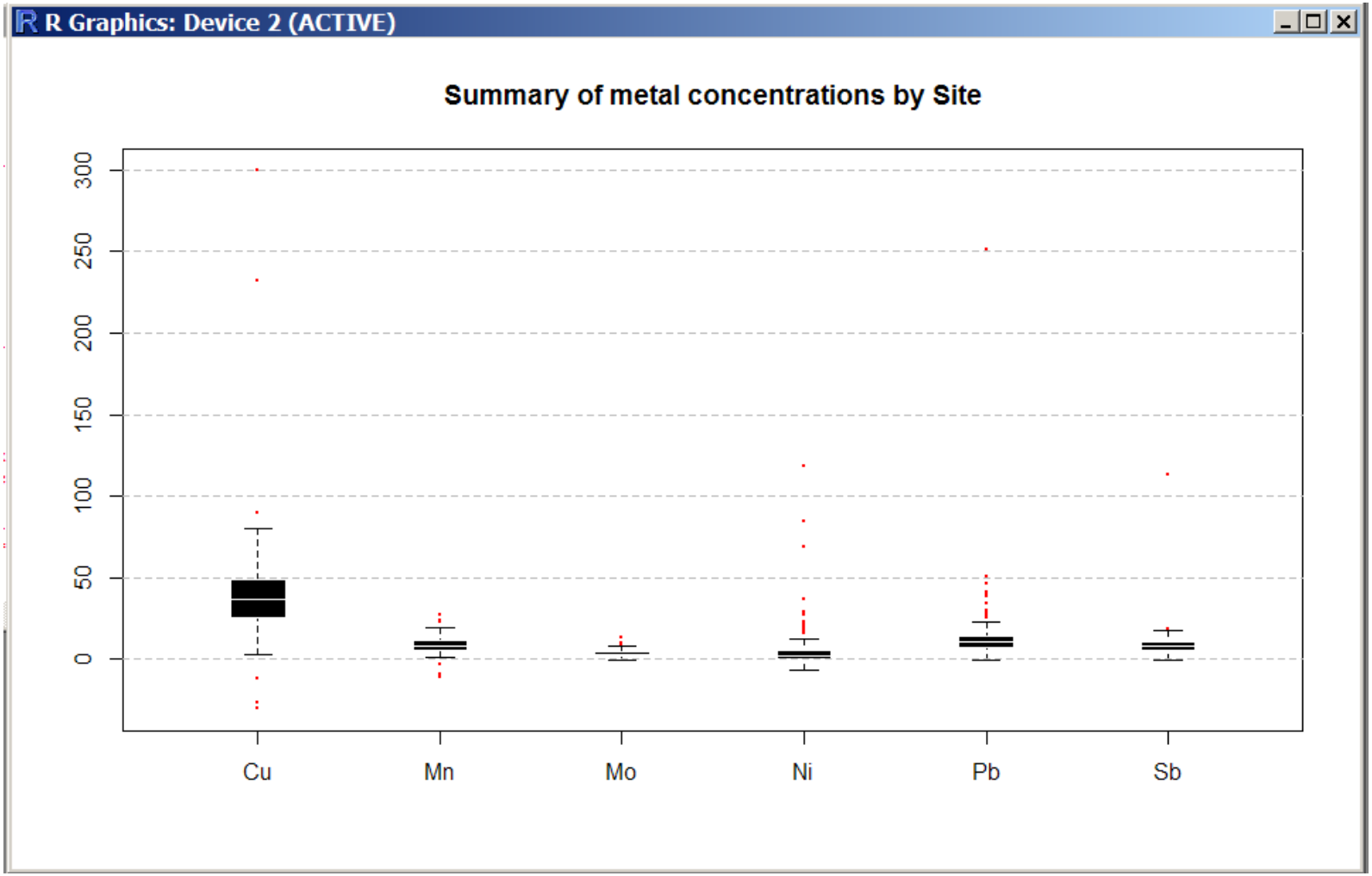
```
boxplot(metals[,-1],
horizontal=TRUE,las=1,
main="Summary of metal
      concentrations by Site")
```



改变箱型风格

```
boxplot(metals[,-1],
border = "white",col = "black",boxwex = 0.3,
medlwd=1, whiskcol="black",staplecol="black",
outcol="red",cex=0.3,outpch=19,
main="Summary of metal concentrations by Site")
grid(nx=NA,ny=NULL,col="gray",lty="dashed")
```

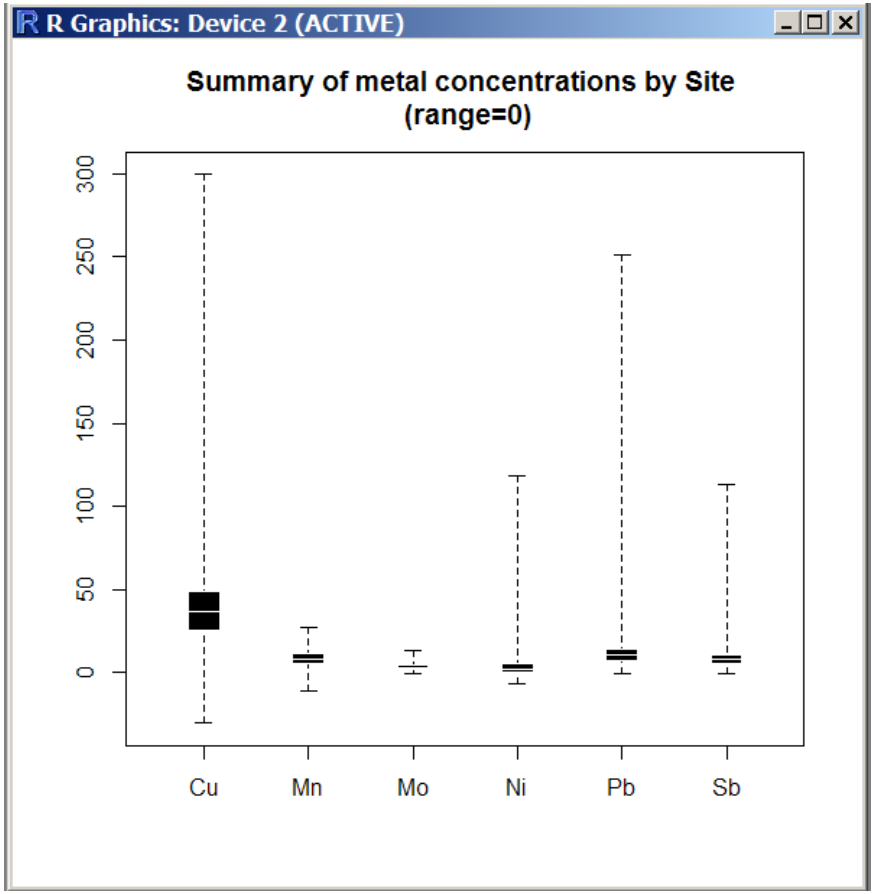
改变箱型风格



2013.04.20

延长须线

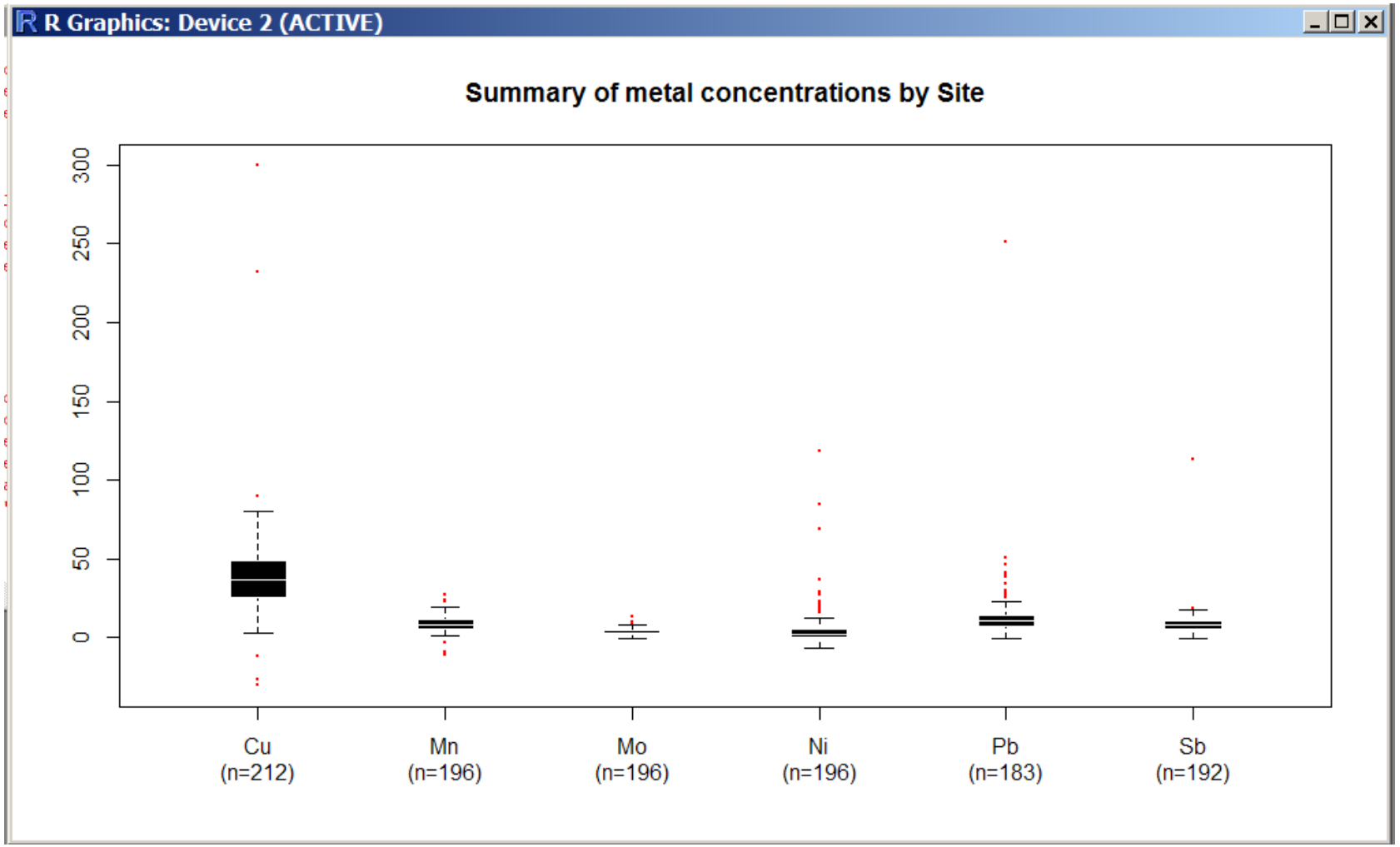
```
boxplot(metals[,-1],
range=0,border = "white",col =
    "black",
boxwex =
    0.3,medlwd=1,whiskcol="black",
staplecol="black",outcol="red",cex=0.
    3,outpch=19,
main="Summary of metal
    concentrations by Site \n
    (range=0)")
```



显示观测数量

```
b<-boxplot(metals[,-1],
xaxt="n",border = "white",col = "black",
boxwex = 0.3,medlwd=1,whiskcol="black",
staplecol="black",outcol="red",cex=0.3,outpch=19,
main="Summary of metal concentrations by Site")
axis(side=1,at=1:length(b$names),
labels=paste(b$names,"\n(n=",b$n,")",sep=""),
mgp=c(3,2,0))
```

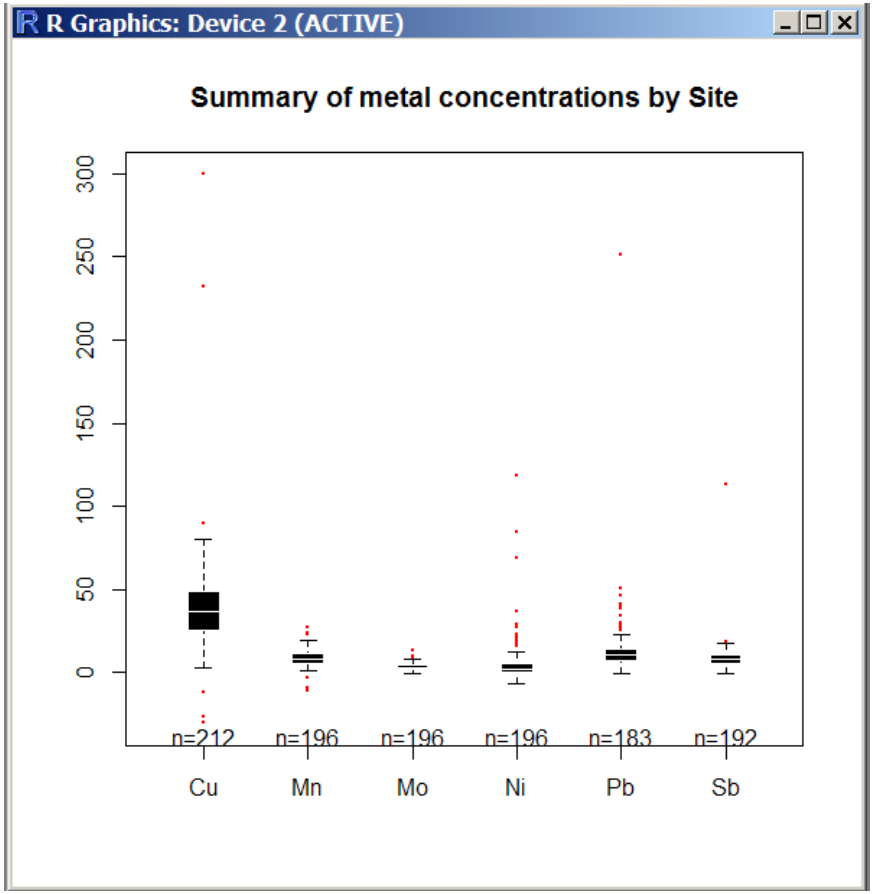
显示观测数量



2013.04.20

使用gplot包

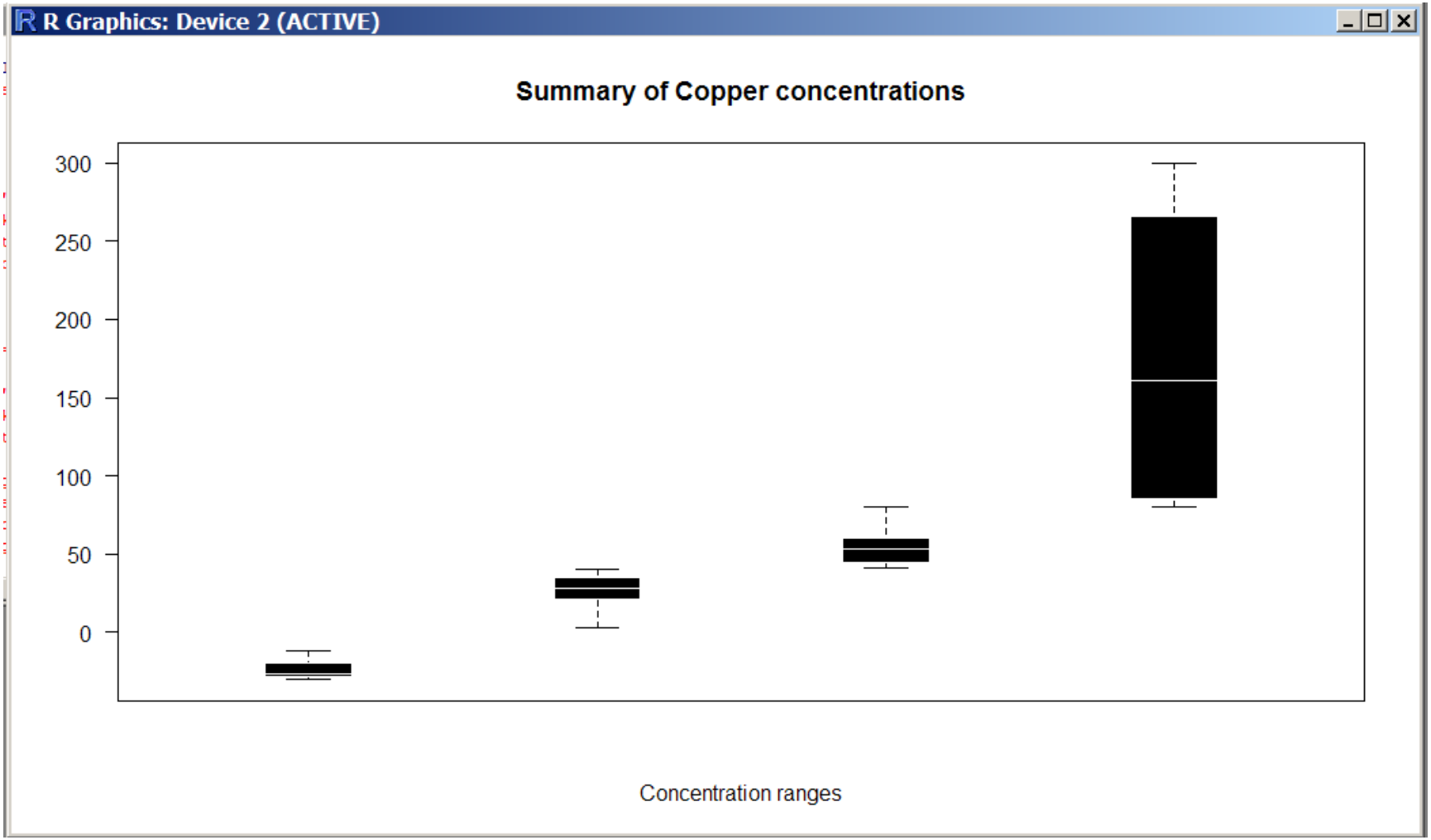
```
install.packages("gplots")
library(gplots)
boxplot.n(metals[,-1],
border = "white",col =
    "black",boxwex = 0.3,
medlwd=1,whiskcol="black",staple
    col="black",
outcol="red",cex=0.3,outpch=19,
main="Summary of metal
    concentrations by Site")
```



分割数据

```
cuts<-c(0,40,80)
Y<-split(x=metals$Cu, f=findInterval(metals$Cu, cuts))
boxplot(Y,xaxt="n",
border = "white",col = "black",boxwex = 0.3,
medlwd=1,whiskcol="black",staplecol="black",
outcol="red",cex=0.3,outpch=19,
main="Summary of Copper concentrations",
xlab="Concentration ranges",las=1)
axis(1,at=1:4,
labels=c("Below 0","0 to 40","40 to 80","Above 80"),
lwd=0,lwd.ticks=1,col="gray")
```

分割数据



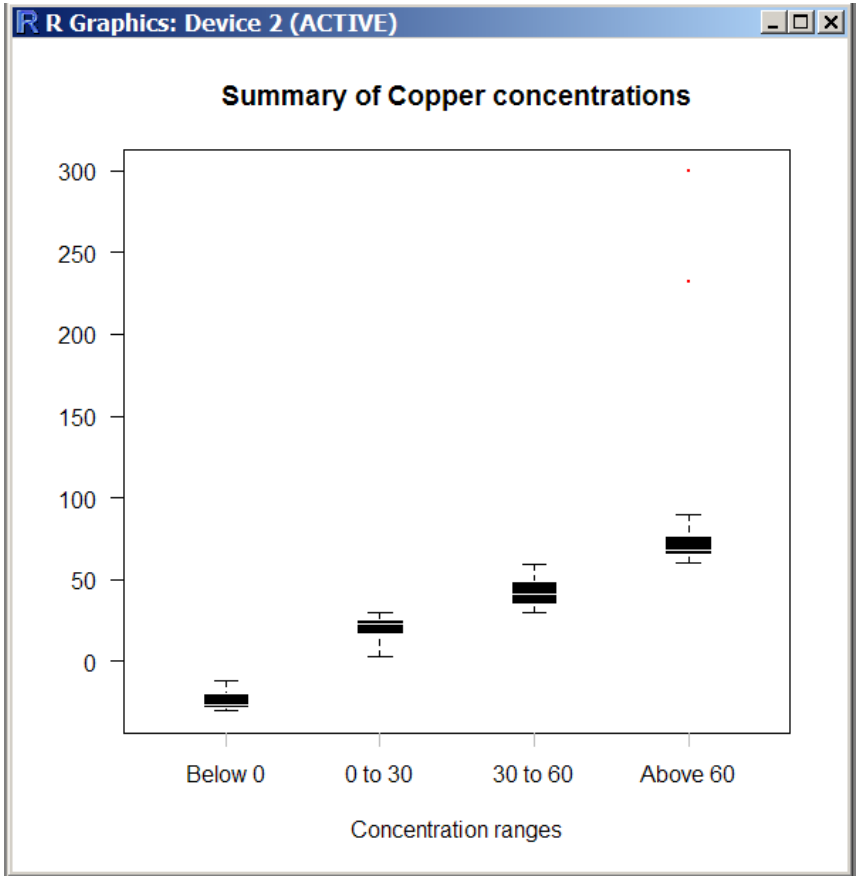
2013.04.20

函数化

```
boxplot.cuts<-function(y,cuts,...) {
Y<-split(metals$Cu, f=findInterval(y, cuts))
b<-boxplot(Y,xaxt="n",
border = "white",col = "black",boxwex = 0.3,
medlwd=1,whiskcol="black",staplecol="black",
outcol="red",cex=0.3,outpch=19,
main="Summary of Copper concentrations",
xlab="Concentration ranges",las=1,...)
clabels<-paste("Below",cuts[1])
for(k in 1:(length(cuts)-1)) {
clabels<-c(clabels, paste(as.character(cuts[k]),
"to", as.character(cuts[k+1])))
}
clabels<-c(clabels,
paste("Above",as.character(cuts[length(cuts)])))
axis(1,at=1:length(clabels),
labels=clabels,lwd=0,lwd.ticks=1,col="gray")
}
```

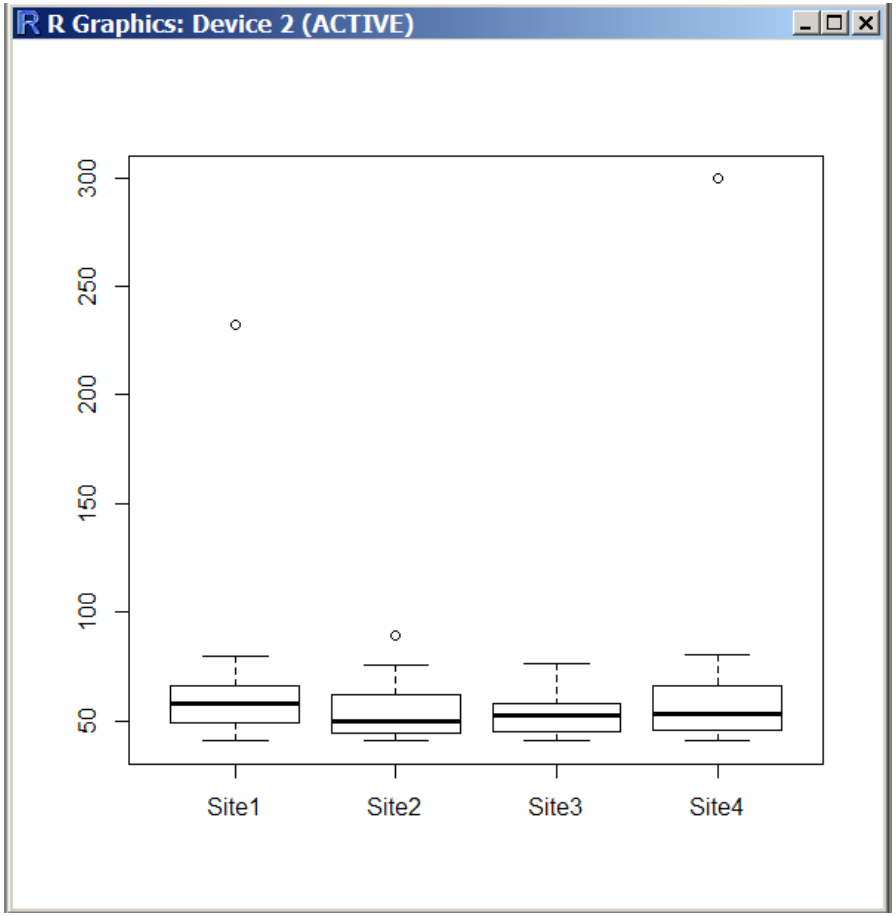
函数化

boxplot.cuts(metals$Cu,c(0,30,60))



子集

boxplot(Cu~Source,data=metals,subset=Cu>40)

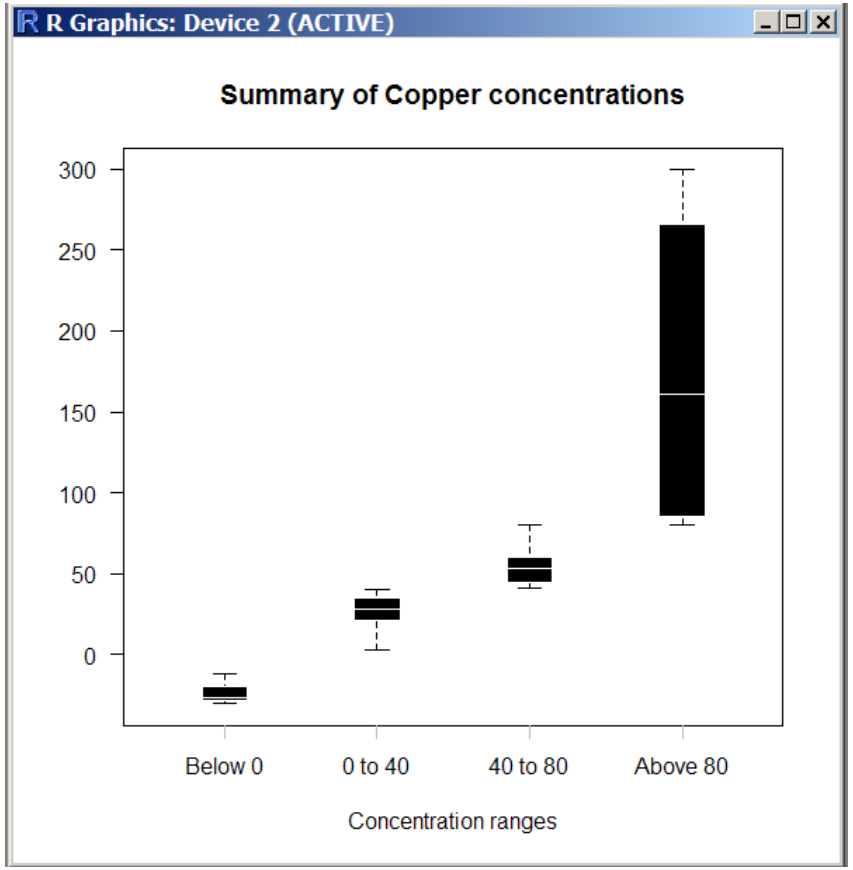


另一个函数

```
boxplot.cuts<-function(y,cuts) {
f=cut(y, c(min(y[!is.na(y)]),cuts,max(y[!is.na(y)])),
ordered_results=TRUE);
Y<-split(y, f=f)
b<-boxplot(Y,xaxt="n",
border = "white",col = "black",boxwex = 0.3,
medlwd=1,whiskcol="black",staplecol="black",
outcol="red",cex=0.3,outpch=19,
main="Summary of Copper concentrations",
xlab="Concentration ranges",las=1)
clabels = as.character(levels(f))
axis(1,at=1:length(clabels),
labels=clabels,lwd=0,lwd.ticks=1,col="gray")
}
```

结果

boxplot.cuts(metals$Cu,c(0,40,80))



热力图

使用第八章数据

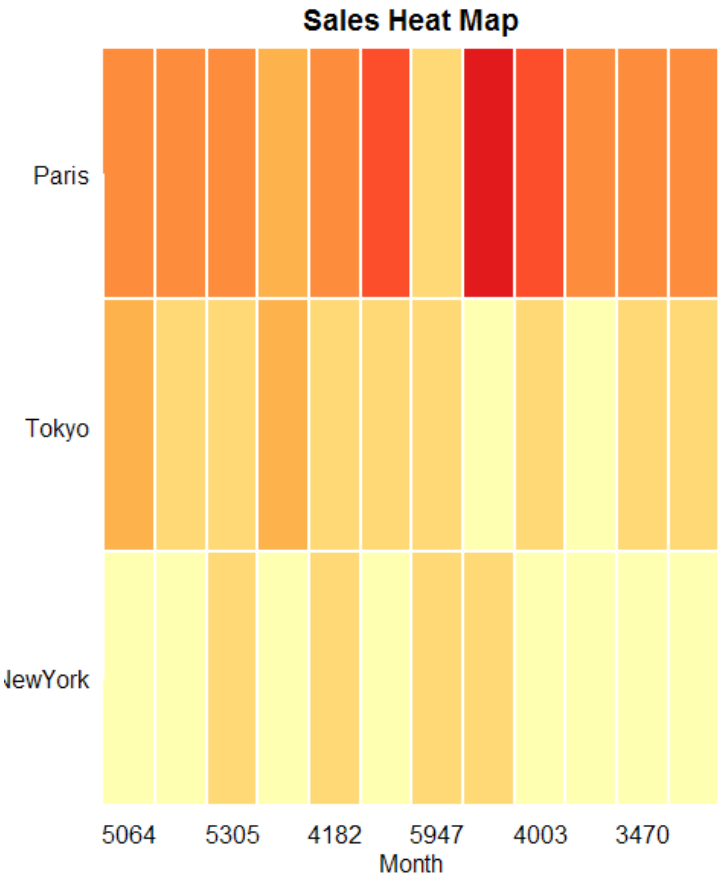
```
sales<-read.csv("sales.csv")


install.packages("RColorBrewer")
library(RColorBrewer)


rownames(sales)<-sales[,1]
sales<-sales[,-1]
data_matrix<-data.mpal=brewer.pal(7,"YlOrRd")
atrix(sales)
breaks<-seq(3000,12000,1500)
```

代码

```
layout(matrix(data=c(1,2), nrow=1, ncol=2), widths=c(8,1),
heights=c(1,1))
#Set margins for the heatmap
par(mar = c(5,10,4,2),oma=c(0.2,0.2,0.2,0.2),mex=0.5)
image(x=1:nrow(data_matrix),y=1:ncol(data_matrix),
z=data_matrix,axes=FALSE,xlab="Month",
ylab="",col=pal[1:(length(breaks)-1)],
breaks=breaks,main="Sales Heat Map")
axis(1,at=1:nrow(data_matrix),labels=rownames(data_matrix),
col="white",las=1)
axis(2,at=1:ncol(data_matrix),labels=colnames(data_matrix),
col="white",las=1)
abline(h=c(1:ncol(data_matrix))+0.5,
v=c(1:nrow(data_matrix))+0.5, col="white",lwd=2,xpd=FALSE)
breaks2<-breaks[-length(breaks)]
```

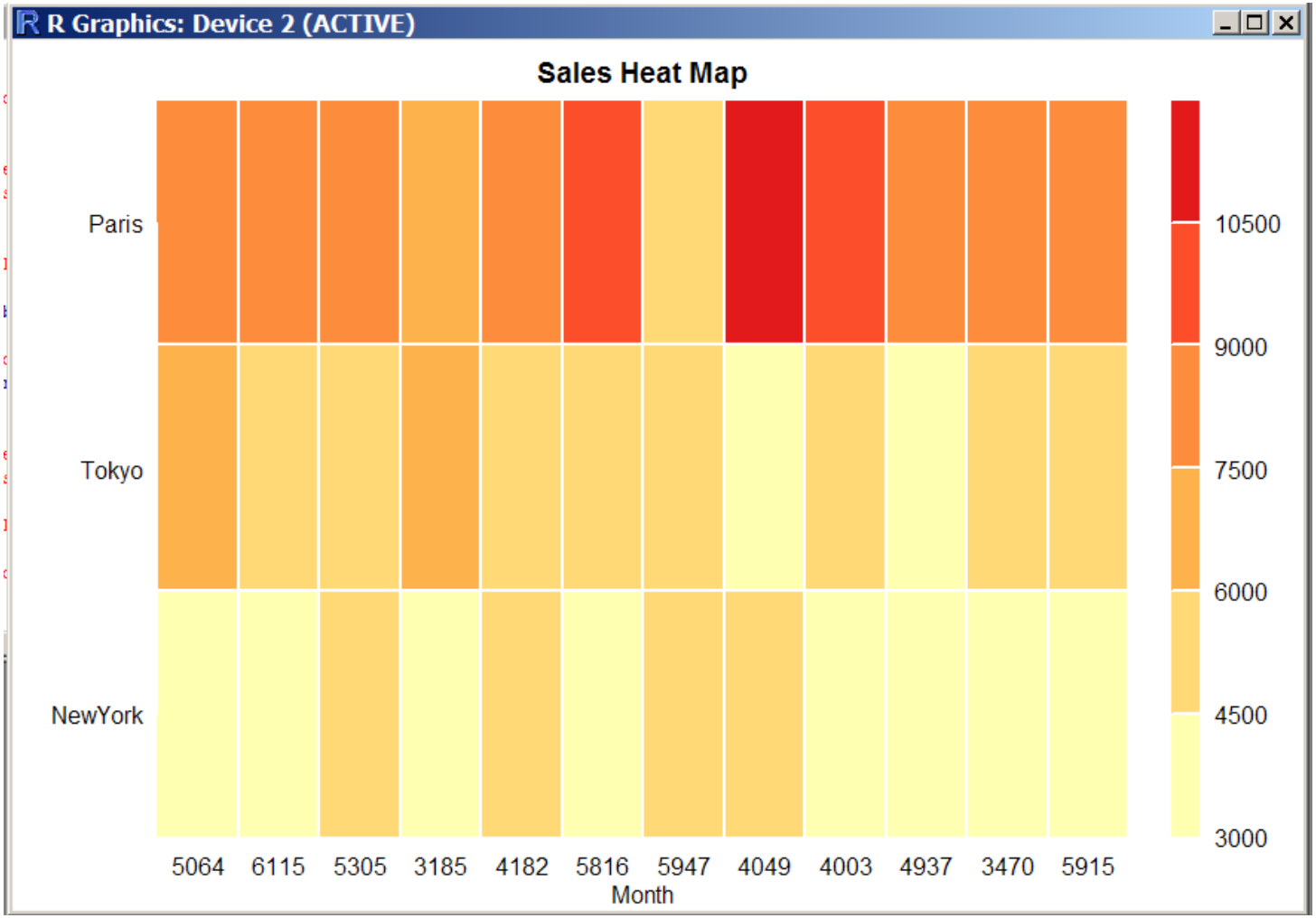


代码

```
par(mar = c(5,1,4,7))


image(x=1, y=0:length(breaks2),z=t(matrix(breaks2))*1.001,
col=pal[1:length(breaks)-1],axes=FALSE,breaks=breaks,
xlab="", ylab="",xaxt="n")
axis(4,at=0:(length(breaks2)-1), labels=breaks2, col="white",
las=1)
abline(h=c(1:length(breaks2)),col="white",lwd=2,xpd=F)
```

结果


R Graphics: Device 2 (ACTIVE)
Sales Heat Map
Paris
Tokyo
NewYork
10500
9000
7500
6000
4500
3000
5064
6115
5305
3185
4182
5816
5947
4049
4003
4937
3470
5915
Month

2013.04.20

相关热力图

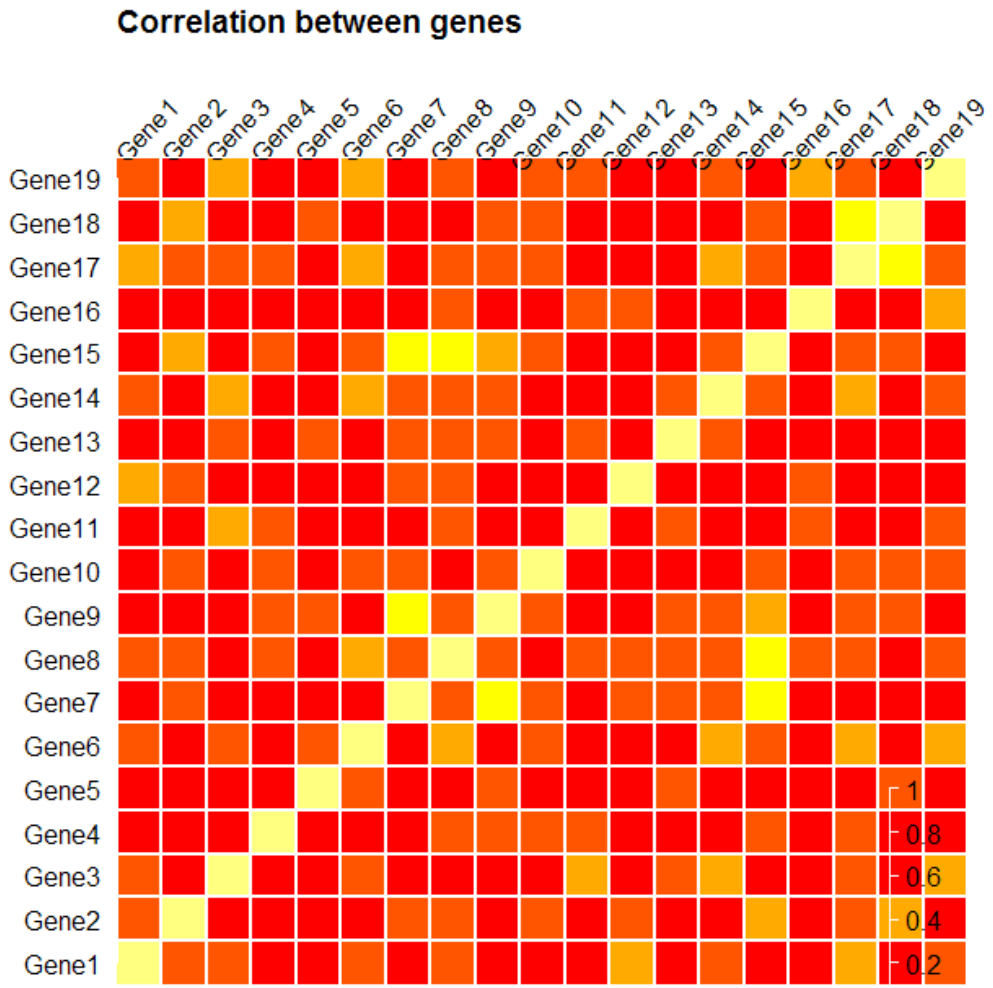
```
genes<-read.csv("genes.csv")


rownames(genes)<-genes[,1]
data_matrix<-data.matrix(genes[,-1])
pal=heat.colors(5)
breaks<-seq(0,1,0.2)
layout(matrix(data=c(1,2), nrow=1, ncol=2), widths=c(8,1),
heights=c(1,1))
par(mar = c(3,7,12,2),oma=c(0.2,0.2,0.2,0.2),mex=0.5)
image(x=1:nrow(data_matrix),y=1:ncol(data_matrix),
z=data_matrix,xlab="",ylab="",breaks=breaks,
col=pal,axes=FALSE)
```

相关热力图

```
text(x=1:nrow(data_matrix)+0.75, y=par("usr")[4] + 1.25,
srt = 45, adj = 1, labels = rownames(data_matrix),
xpd = TRUE)
axis(2,at=1:ncol(data_matrix),labels=colnames(data_matrix),
col="white",las=1)
abline(h=c(1:ncol(data_matrix))+0.5,v=c(1:nrow(data_matrix))+0.5,
col="white",lwd=2,xpd=F)
title("Correlation between genes",line=8,adj=0)
breaks2<-breaks[-length(breaks)]
# Color Scale
par(mar = c(25,1,25,7))
image(x=1, y=0:length(breaks2),z=t(matrix(breaks2))*1.001,
col=pal[1:length(breaks)-1],axes=FALSE,
breaks=breaks,xlab="",ylab="",
xaxt="n")
axis(4,at=0:(length(breaks2)),labels=breaks,col="white",las=1)
abline(h=c(1:length(breaks2)),col="white",lwd=2,xpd=F)
```

相关热力图



2013.04.20

展现多变量数据

```
rownames(nba)<-nba[,1]
data_matrix<-t(scale(data.matrix(nba[,-1])))
pal=brewer.pal(6,"Blues")
statnames<-c("Games Played", "Minutes Played", "Total Points",
"Field Goals Made", "Field Goals Attempted",
"Field Goal Percentage", "Free Throws Made",
"Free Throws Attempted", "Free Throw Percentage",
"Three Pointers Made", "Three Pointers Attempted",
"Three Point Percentage", "Offensive Rebounds",
"Defensive Rebounds", "Total Rebounds", "Assists", "Steals",
"Blocks", "Turnovers", "Fouls")
par(mar = c(3,14,19,2),oma=c(0.2,0.2,0.2,0.2),mex=0.5)
#Heat map
image(x=1:nrow(data_matrix),y=1:ncol(data_matrix),
z=data_matrix,xlab="",ylab="",col=pal,axes=FALSE)
```

展现多变量数据

```
#X axis labels
text(1:nrow(data_matrix), par("usr")[4] + 1,
srt = 45, adj = 0,labels = statnames,
xpd = TRUE, cex=0.85)
#Y axis labels
axis(side=2,at=1:ncol(data_matrix),
labels=colnames(data_matrix),
col="white",las=1, cex.axis=0.85)
#White separating lines
abline(h=c(1:ncol(data_matrix))+0.5,
v=c(1:nrow(data_matrix))+0.5,
col="white",lwd=1,xpd=F)
#Graph Title
text(par("usr")[1]+5, par("usr")[4] + 12,
"NBA per game performance of top 50corers",
xpd=TRUE,font=2,cex=1.5)
```

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间