海藻数据的分析实验报告

1、把从网上下载的 Analysis.txt 文件中的数据读入到 R 中：

```
algae<-read.table('Analysis.txt',
       header=F, dec='.', col.names=c('season','size','speed','mxPH','mnO2','Cl',
      'NO3','NH4','oPO4','PO4','Chla',
      'a1','a2','a3','a4','a5','a6','a7'), na.strings=c('XXXXXXX'))
```

2、数据可视化和摘要

(1) summary(algae)　# 获取数据的如下描述性统计摘要：

```
> summary(algae)
    season       size       speed        mxPH            mnO2
 autumn:40   large :45   high  :84   Min.   :5.600   Min.   : 1.500
 spring:53   medium:84   low   :33   1st Qu.:7.700   1st Qu.: 7.725
 summer:45   small :71   medium:83   Median :8.060   Median : 9.800
 winter:62                           Mean   :8.012   Mean   : 9.118
                                     3rd Qu.:8.400   3rd Qu.:10.800
                                     Max.   :9.700   Max.   :13.400
                                     NA's   :1       NA's   :2
       Cl                NO3               NH4                oPO4
 Min.   :  0.222   Min.   : 0.050   Min.   :    5.00   Min.   :  1.00
 1st Qu.: 10.981   1st Qu.: 1.296   1st Qu.:   38.33   1st Qu.: 15.70
 Median : 32.730   Median : 2.675   Median :  103.17   Median : 40.15
 Mean   : 43.636   Mean   : 3.282   Mean   :  501.30   Mean   : 73.59
 3rd Qu.: 57.824   3rd Qu.: 4.446   3rd Qu.:  226.95   3rd Qu.: 99.33
 Max.   :391.500   Max.   :45.650   Max.   :24064.00   Max.   :564.60
 NA's   :10        NA's   :2        NA's   :2          NA's   :2
      PO4               Chla              a1               a2
 Min.   :  1.00   Min.   :  0.200   Min.   : 0.00   Min.   : 0.000
 1st Qu.: 41.38   1st Qu.:  2.000   1st Qu.: 1.50   1st Qu.: 0.000
 Median :103.29   Median :  5.475   Median : 6.95   Median : 3.000
 Mean   :137.88   Mean   : 13.971   Mean   :16.92   Mean   : 7.458
 3rd Qu.:213.75   3rd Qu.: 18.308   3rd Qu.:24.80   3rd Qu.:11.375
 Max.   :771.60   Max.   :110.456   Max.   :89.80   Max.   :72.600
 NA's   :2        NA's   :12
       a3               a4               a5               a6
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
 Median : 1.550   Median : 0.000   Median : 1.900   Median : 0.000
 Mean   : 4.309   Mean   : 1.992   Mean   : 5.064   Mean   : 5.964
 3rd Qu.: 4.925   3rd Qu.: 2.400   3rd Qu.: 7.500   3rd Qu.: 6.925
 Max.   :42.800   Max.   :44.600   Max.   :44.400   Max.   :77.600

       a7
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 1.000
 Mean   : 2.495
 3rd Qu.: 2.400
 Max.   :31.600
```

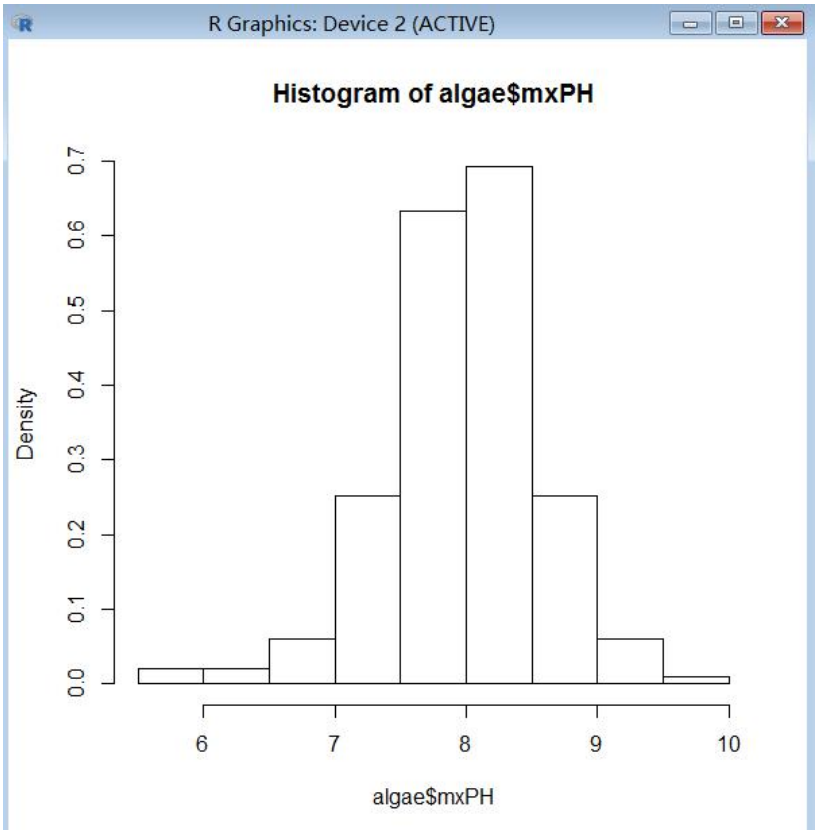
(2)列出了均值、中位数、四分位数及极值等统计信息。

提供了变量值分布的初步信息。NA 后面的数字表示缺失值的个数。通过观

察数据的统计特性我们可以了解数据分布的偏度和分散情况。

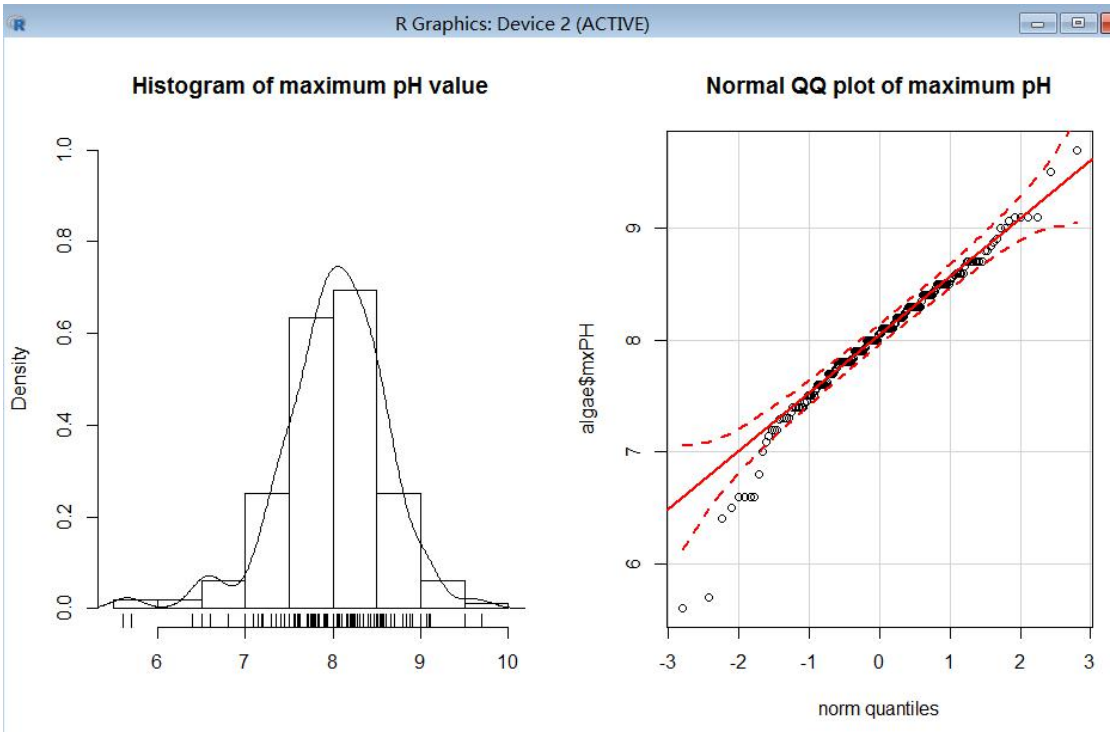
(3)绘制变量 mxPH 的直方图：

hist(algae$mxPH,prob = T)　# 设置参数 prob = T 可以得到每个取值区间的概率，若设为 false，则给出频数。从下图中可以看出变量 mxPH 的分布非常接近正态分布，它的值大部分聚集在该变量的均值周围。



用 qq 图检验其分布是否为正态分布。获取 Q-Q 图的指令如下：

```
library(car)          # 载入 R 的添加包 car
par(mfrow=c(1,2))    # 把图形输出窗口设置为 1 行 2 列
hist(algae$mxPH,prob=T,xlab='',    # 绘制变量 mxPH 的直方图
main='Histogram of maximum pH value',ylim=0:1)    # 设置 y 轴范围
lines(density(algae$mxPH,na.rm=T))    # 绘制平滑版本的直方图(变量分布的核密度估计)
rug(jitter(algae$mxPH))    # 在 x 轴附近绘制变量的实际值，易于识别离群点
qq.plot(algae$mxPH,main='Normal QQ plot of maximum pH')
```

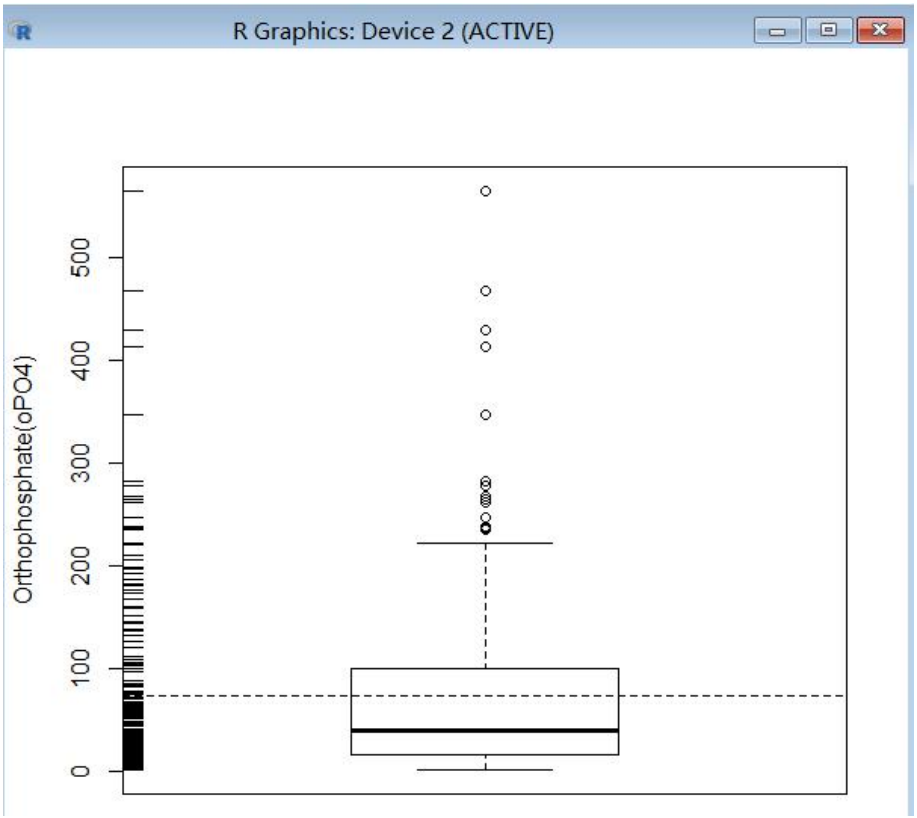


左图是一个复杂的直方图，观察到有两个值明显低于所有其他值，很可能是离群点，也即数据样本中可能出现的错误，可以帮助定位奇怪的错误值并在后续分析中进行剔除。

右面即是用函数 qq.plot()函数得到的 Q-Q 图。绘制变量值和正态分布的理论分位数的散点图。给出了正态分布的 95%置信区间的带状图。观察到变量有几个小的值明显在 95%置信区间之外，它们不服从正态分布。

(4)绘制盒图，对离群值进行识别：

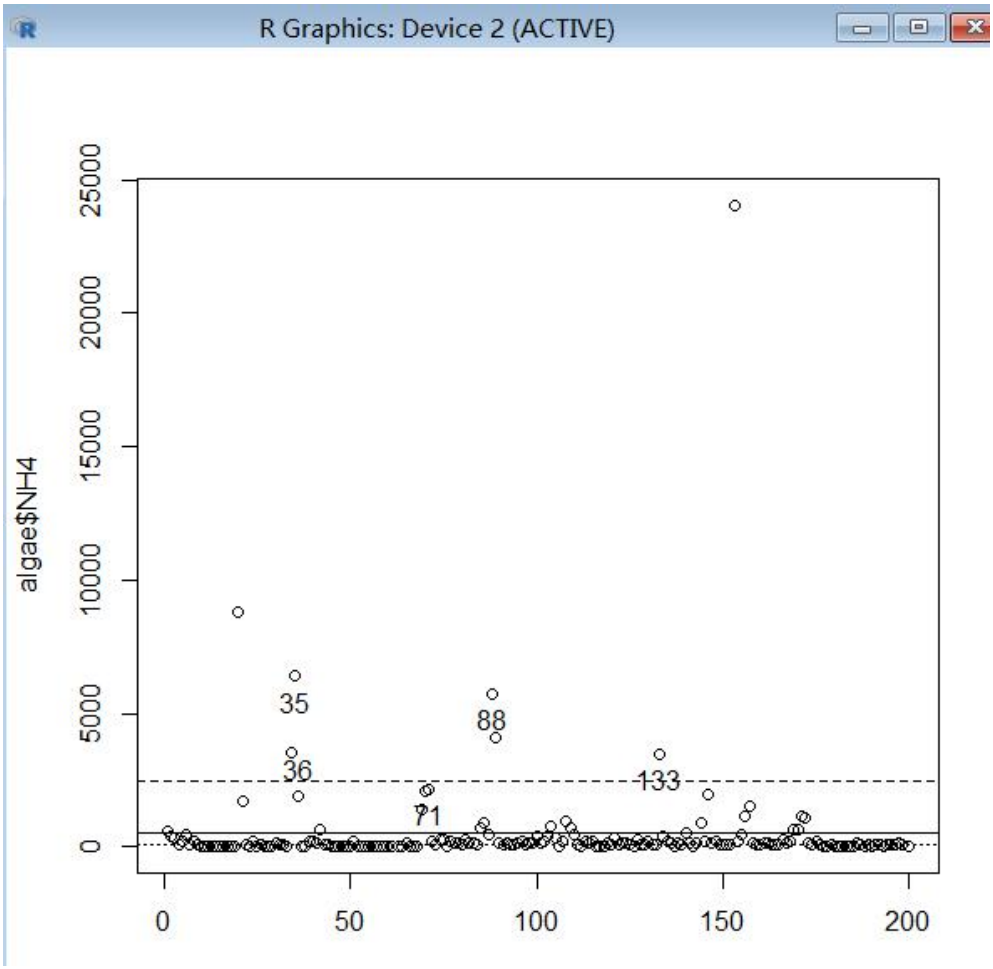
```
boxplot(algae$oPO4,ylab="Orthophosphate(oPO4)")  # 绘制变量 oPO4 的盒图
rug(jitter(algae$oPO4),side = 2)
abline(h=mean(algae$oPO4,na.rm=T),lty=2)  # 使用 abline 在变量的均值位置绘制一条水平线，mean()函数用来计算均值
```



盒图框的边界代表变量的第一个四分位数和第三个四分位数，框内的水平线是变量的中位数。设 r 是变量的四分位距，盒图上方的小横线是<=第三个四分位数+1.5*r 的最大的观测值，而下方的小横线是>=第一个四分位数-1.5*r 的最小观测值。盒图上方的小横线上面的小圆圈表示与其他值相比特别大的值，通常认为是离群值。盒图不仅给出了变量的中心趋势，也给出了变量的发散情况和离群值。

变量 oPO4 的分布集中在较小的观测值周围，因此分布为正偏。大部分水样的 oPO4 值比较低，但也有几个水样的观测值较高。确定有离群值的观测。用下列方式识别特大值相应的水样：

```
plot(algae$NH4,xlab='')     # 绘制变量的所有值
abline(h=mean(algae$NH4,na.rm=T),lty=1)       # 均值
abline(h=mean(algae$NH4,na.rm=T)+sd(algae$NH4,na.rm=T),
lty=2)   # 均值加一个标准差
abline(h=median(algae$NH4,na.rm=T),lty=3)   # 中位数
identify(algae$NH4)       # 交互式指令，允许用户单击图形中的点
```



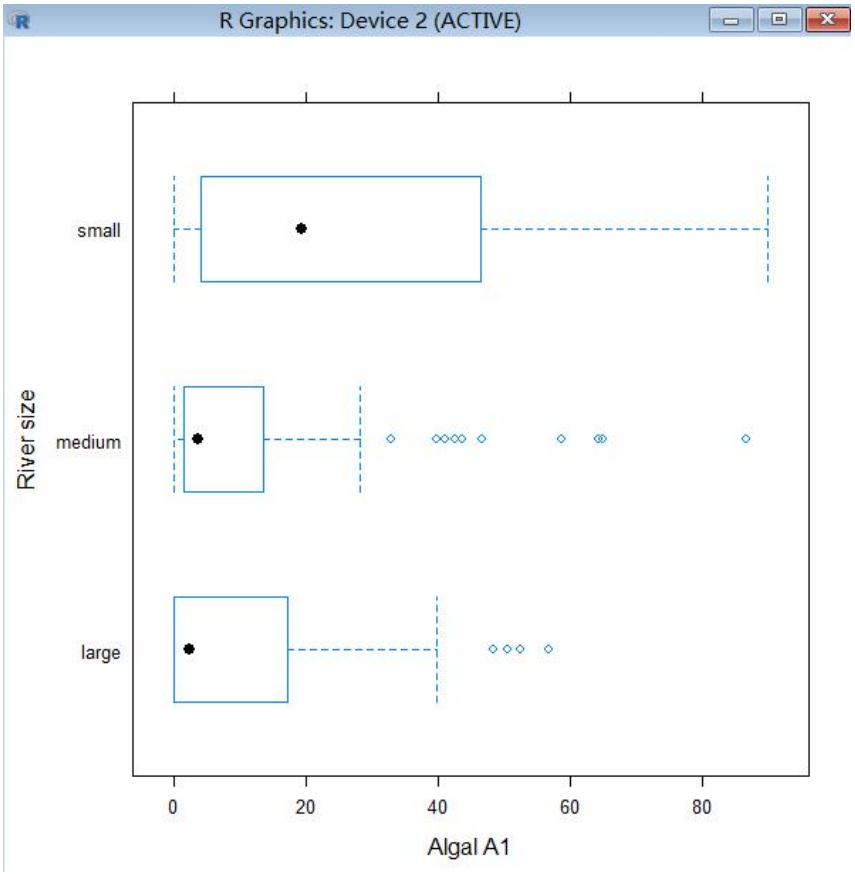
研究海藻变量 a1 值的分布。对于变量 size 的不同取值，可以绘制变量 a1 的一组盒图，每个盒图对应于变量 size 的某个特定值的水样子集。

研究 size 如何影响 a1 的值分布：

```
library(lattice)      # 载入 lattice 包
bwplot(size ~ a1,data=algae,ylab='River size',xlab='Algal A1') # 绘制这些图
```

lattice 版本的盒图对变量 size 的每个值绘制 a1，同理 a2-a7。

下面给出海藻变量 a1 的条件箱图，从中可以看出，规模小的河流，海藻的频率较高。



3、数据缺失的处理　——四种策略

(1)将缺失部分剔除

检测某些变量中至少含有一个缺失数据的所有观测值，得到这些观测值的个数：algae[!complete.cases(algae),]　　# complete.cases()产生一个布尔值向量，该向量的元素个数与 algae 数据框中的行数相同，若数据框中相应行中不含 NA 值（即为一个完整的观测值），函数返回值就是 TRUE。

```
> algae[!complete.cases(algae),]
    season   size  speed mxPH mnO2     Cl    NO3 NH4    oPO4     PO4  Chla    a1   a2  a3   a4  a5  a6  a7
28  autumn  small   high 6.80 11.1 9.000 0.630  20   4.000      NA  2.70 30.3  1.9 0.0  0.0 2.1 1.4 2.1
38  spring  small   high 8.00   NA 1.450 0.810  10   2.500   3.000  0.30 75.8  0.0 0.0  0.0 0.0 0.0 0.0
48  winter  small    low   NA 12.6 9.000 0.230  10   5.000   6.000  1.10 35.5  0.0 0.0  0.0 0.0 0.0 0.0
55  winter  small   high 6.60 10.8    NA 3.245  10   1.000   6.500    NA 24.3  0.0 0.0  0.0 0.0 0.0 0.0
56  spring  small medium 5.60 11.8    NA 2.220   5   1.000   1.000    NA 82.7  0.0 0.0  0.0 0.0 0.0 0.0
57  autumn  small medium 5.70 10.8    NA 2.550  10   1.000   4.000    NA 16.8  4.6 3.9 11.5 0.0 0.0 0.0
58  spring  small   high 6.60  9.5    NA 1.320  20   1.000   6.000    NA 46.8  0.0 0.0 28.8 0.0 0.0 0.0
59  summer  small   high 6.60 10.8    NA 2.640  10   2.000  11.000    NA 46.9  0.0 0.0 13.4 0.0 0.0 0.0
60  autumn  small medium 6.60 11.3    NA 4.170  10   1.000   6.000    NA 47.1  0.0 0.0  0.0 0.0 1.2 0.0
61  spring  small medium 6.50 10.4    NA 5.970  10   2.000  14.000    NA 66.9  0.0 0.0  0.0 0.0 0.0 0.0
62  summer  small medium 6.40   NA    NA    NA  NA      NA  14.000    NA 19.4  0.0 0.0  2.0 0.0 3.9 1.7
63  autumn  small   high 7.83 11.7 4.083 1.328  18   3.333   6.667    NA 14.4  0.0 0.0  0.0 0.0 0.0 0.0
116 winter medium   high 9.70 10.8 0.222 0.406  10  22.444  10.111    NA 41.0  1.5 0.0  0.0 0.0 0.0 0.0
161 spring  large    low 9.00  5.8    NA 0.900 142 102.000 186.000 68.05  1.7 20.6 1.5  2.2 0.0 0.0 0.0
184 winter  large   high 8.00 10.9 9.055 0.825  40  21.083  56.091    NA 16.8 19.6 4.0  0.0 0.0 0.0 0.0
199 winter  large medium 8.00  7.6    NA    NA  NA      NA      NA    NA  0.0 12.5 3.7  1.0 0.0 0.0 4.9
```

nrow(algae[!complete.cases(algae),])

[1] 16　　# 共有 16 条记录有缺失现象

```
> algae[!complete.cases(algae),]
    season   size  speed mxPH mnO2    Cl   NO3 NH4    oPO4      PO4  Chla   a1   a2  a3
28  autumn  small   high 6.80 11.1 9.000 0.630  20   4.000       NA  2.70 30.3  1.9 0.0
38  spring  small   high 8.00   NA 1.450 0.810  10   2.500    3.000  0.30 75.8  0.0 0.0
48  winter  small    low   NA 12.6 9.000 0.230  10   5.000    6.000  1.10 35.5  0.0 0.0
55  winter  small   high 6.60 10.8    NA 3.245  10   1.000    6.500    NA 24.3  0.0 0.0
56  spring  small medium 5.60 11.8    NA 2.220   5   1.000    1.000    NA 82.7  0.0 0.0
57  autumn  small medium 5.70 10.8    NA 2.550  10   1.000    4.000    NA 16.8  4.6 3.9
58  spring  small   high 6.60  9.5    NA 1.320  20   1.000    6.000    NA 46.8  0.0 0.0
59  summer  small   high 6.60 10.8    NA 2.640  10   2.000   11.000    NA 46.9  0.0 0.0
60  autumn  small medium 6.60 11.3    NA 4.170  10   1.000    6.000    NA 47.1  0.0 0.0
61  spring  small medium 6.50 10.4    NA 5.970  10   2.000   14.000    NA 66.9  0.0 0.0
62  summer  small medium 6.40   NA    NA    NA  NA      NA   14.000    NA 19.4  0.0 0.0
63  autumn  small   high 7.83 11.7 4.083 1.328  18   3.333    6.667    NA 14.4  0.0 0.0
116 winter medium   high 9.70 10.8 0.222 0.406  10  22.444   10.111    NA 41.0  1.5 0.0
161 spring  large    low 9.00  5.8    NA 0.900 142 102.000  186.000 68.05  1.7 20.6 1.5
184 winter  large   high 8.00 10.9 9.055 0.825  40  21.083   56.091    NA 16.8 19.6 4.0
199 winter  large medium 8.00  7.6    NA    NA  NA      NA       NA    NA  0.0 12.5 3.7
      a4  a5  a6  a7
28   0.0 2.1 1.4 2.1
38   0.0 0.0 0.0 0.0
48   0.0 0.0 0.0 0.0
55   0.0 0.0 0.0 0.0
56   0.0 0.0 0.0 0.0
57  11.5 0.0 0.0 0.0
58  28.8 0.0 0.0 0.0
59  13.4 0.0 0.0 0.0
60   0.0 0.0 1.2 0.0
61   0.0 0.0 0.0 0.0
62   2.0 0.0 3.9 1.7
63   0.0 0.0 0.0 0.0
116  0.0 0.0 0.0 0.0
161  2.2 0.0 0.0 0.0
184  0.0 0.0 0.0 0.0
199  1.0 0.0 0.0 4.9
```

algae<-na.omit(algae)　# 从数据框中剔除这 16 个样本

nrow(algae[!complete.cases(algae),])　# 再次检查含有缺失值的水样记录

```
> nrow(algae[!complete.cases(algae),])
[1] 0
>
```

显示已经为 0，说明我们已经把含有缺失值的水样剔除掉了

观察到这个样本中数据第 62 条和第 199 条记录中的 11 个解释变量中有 6 个是缺失值，这种情况剔除掉它们：

algae<-algae[-c(62,199),]

找出海藻数据集中每行数据的缺失值个数：

apply(algae, 1, function(x) sum(is.na(x)))

```
> apply(algae, 1, function(x) sum(is.na(x)))
  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
 [37] 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 2 2 2 2 2 2 2 6 1 0 0 0 0 0 0 0 0
 [73] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[109] 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[145] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[181] 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 6 0
>
```

(2) 用最高频率值来填补缺失值

	season	size	speed	mxPH	mnO2	Cl	NO3	NH4	oPO4	PO4	Chla	a1	a2	a3
48	winter	small	low	NA	12.60	9.000	0.230	10.000	5.000	6.000	1.100	35.5	0.0	0.0

例如样本 algae[48,]中的变量 mxPH 有缺失值(上图显示 NA)，由于该变量分布近似正态分布，因此我们选用平均值来填补这个空缺：

algae[48,"mxPH"]<-mean(algae$mxPH,na.rm=T) # mean()计算数值向量的平均值，参数 na.rm=T 使计算时忽略缺失数据。

48	winter	small	low	8.011734	12.60	9.000	0.230	10.000	5.000	6.000	1.100

例如 Chla，这个变量在 12 行上有缺失值，它的分布偏向于较低的数值，并且有几个极端值，因此使用中位数来填补这一类的缺失值：

algae[is.na(algae$Chla),"Chla"]<-median(algae$Chla,na.rm=T)

首先显示含有缺失值的记录： algae[!complete.cases(algae),]

```
> algae[!complete.cases(algae),]
    season   size  speed mxPH mnO2     Cl   NO3 NH4    oPO4     PO4  Chla   a1   a2  a3   a4  a5  a6  a7
28  autumn  small   high 6.80 11.1 9.000 0.630  20   4.000      NA  2.70 30.3  1.9 0.0  0.0 2.1 1.4 2.1
38  spring  small   high 8.00   NA 1.450 0.810  10   2.500   3.000  0.30 75.8  0.0 0.0  0.0 0.0 0.0 0.0
48  winter  small    low   NA 12.6 9.000 0.230  10   5.000   6.000  1.10 35.5  0.0 0.0  0.0 0.0 0.0 0.0
55  winter  small   high 6.60 10.8    NA 3.245  10   1.000   6.500    NA 24.3  0.0 0.0  0.0 0.0 0.0 0.0
56  spring  small medium 5.60 11.8    NA 2.220   5   1.000   1.000    NA 82.7  0.0 0.0  0.0 0.0 0.0 0.0
57  autumn  small medium 5.70 10.8    NA 2.550  10   1.000   4.000    NA 16.8  4.6 3.9 11.5 0.0 0.0 0.0
58  spring  small   high 6.60  9.5    NA 1.320  20   1.000   6.000    NA 46.8  0.0 0.0 28.8 0.0 0.0 0.0
59  summer  small   high 6.60 10.8    NA 2.640  10   2.000  11.000    NA 46.9  0.0 0.0 13.4 0.0 0.0 0.0
60  autumn  small medium 6.60 11.3    NA 4.170  10   1.000   6.000    NA 47.1  0.0 0.0  0.0 0.0 1.2 0.0
61  spring  small medium 6.50 10.4    NA 5.970  10   2.000  14.000    NA 66.9  0.0 0.0  0.0 0.0 0.0 0.0
62  summer  small medium 6.40   NA    NA    NA  NA      NA  14.000    NA 19.4  0.0 0.0  2.0 0.0 3.9 1.7
63  autumn  small   high 7.83 11.7 4.083 1.328  18   3.333   6.667    NA 14.4  0.0 0.0  0.0 0.0 0.0 0.0
116 winter medium   high 9.70 10.8 0.222 0.406  10  22.444  10.111    NA 41.0  1.5 0.0  0.0 0.0 0.0 0.0
161 spring  large    low 9.00  5.8    NA 0.900 142 102.000 186.000 68.05  1.7 20.6 1.5  2.2 0.0 0.0 0.0
184 winter  large   high 8.00 10.9 9.055 0.825  40  21.083  56.091    NA 16.8 19.6 4.0  0.0 0.0 0.0 0.0
199 winter  large medium 8.00  7.6    NA    NA  NA      NA      NA    NA  0.0 12.5 3.7  1.0 0.0 0.0 4.9
```

使用函数 centralImputation()可以用数据的中心趋势值来填补数据集的所有缺失值。对数值型变量，该函数用中位数；对名义变量，它采用众数。

```
> library(DMwR)
载入需要的程辑包：lattice
载入需要的程辑包：grid
> data(algae)
> algae<-algae[-manyNAs(algae),]
> algae<-centralImputation(algae)
> algae[!complete.cases(algae),]
 [1] season size   speed  mxPH   mnO2   Cl     NO3    NH4    oPO4   PO4
[11] Chla   a1     a2     a3     a4     a5     a6     a7
<0 行> (或0-长度的row.names)
>
```

最后再用 algae[!complete.cases(algae),]检验是否含有缺失值的记录，可以看出已经为 0 行.

(3)通过属性的相关关系来填补缺失值

通过变量值某变量与 mxPH 高度相关。使我们得到含有缺失值的第 48 条样本的更可能的填补值。得到变量间的相关值：cor(algae[,4:18],use="complete.obs")

```
> cor(algae[,4:18],use="complete.obs")
             mxPH        mnO2          Cl         NO3         NH4         oPO4         PO4        Chla          a1
mxPH  1.00000000 -0.10025795  0.14602737 -0.17100671 -0.15567926  0.086554432  0.09743184  0.43140811 -0.16728010
mnO2 -0.10025795  1.00000000 -0.26324536  0.11790769 -0.07826816 -0.393752688 -0.46396073 -0.13121671  0.24998372
Cl    0.14602737 -0.26324536  1.00000000  0.21095831  0.06598336  0.379255958  0.44519118  0.14295776 -0.35923946
NO3  -0.17100671  0.11790769  0.21095831  1.00000000  0.72467766  0.133014517  0.15702971  0.14549290 -0.24723921
NH4  -0.15567926 -0.07826816  0.06598336  0.72467766  1.00000000  0.219311206  0.19939575  0.09120406 -0.12360578
oPO4  0.08655443 -0.39375269  0.37925596  0.13301452  0.21931121  1.000000000  0.91196460  0.10691478 -0.39457448
PO4   0.09743184 -0.46396073  0.44519118  0.15702971  0.19939575  0.911964602  1.00000000  0.24849223 -0.45816781
Chla  0.43140811 -0.13121671  0.14295776  0.14549290  0.09120406  0.106914784  0.24849223  1.00000000 -0.26601088
a1   -0.16728010  0.24998372 -0.35923946 -0.24723921 -0.12360578 -0.394574479 -0.45816781 -0.26601088  1.00000000
a2    0.33213028 -0.06848199  0.07845402  0.01997079 -0.03790296  0.123811068  0.13266789  0.36672465 -0.26266549
a3   -0.03082144 -0.23522831  0.07653027 -0.09182236 -0.11290467  0.005704557  0.03219398 -0.06330113 -0.10817758
a4   -0.18103138 -0.37982999  0.14147281 -0.01448875  0.27452000  0.382481433  0.40883951 -0.08600540 -0.09338072
a5   -0.11073064  0.21001174  0.14534877  0.21213579  0.01544458  0.122027482  0.15548900 -0.07342837 -0.26972709
a6   -0.17620457  0.18862656  0.16904394  0.54404455  0.40119275  0.003340366  0.05320294  0.01032550 -0.26156402
a7   -0.16946706 -0.10455106 -0.04494524  0.07505030 -0.02539279  0.026150420  0.07978353  0.01760782 -0.19306384
             a2          a3          a4          a5          a6          a7
mxPH  0.332130278 -0.030821443 -0.18103138 -0.11073064 -0.176204571 -0.16946706
mnO2 -0.068481989 -0.235228307 -0.37982999  0.21001174  0.188626555 -0.10455106
Cl    0.078454019  0.076530269  0.14147281  0.14534877  0.169043945 -0.04494524
NO3   0.019970786 -0.091822358 -0.01448875  0.21213579  0.544044553  0.07505030
NH4  -0.037902958 -0.112904666  0.27452000  0.01544458  0.401192749 -0.02539279
oPO4  0.123811068  0.005704557  0.38248143  0.12202748  0.003340366  0.02615042
PO4   0.132667891  0.032193981  0.40883951  0.15548900  0.053202942  0.07978353
Chla  0.366724647 -0.063301128 -0.08600540 -0.07342837  0.010325497  0.01760782
a1   -0.262665485 -0.108177581 -0.09338072 -0.26972709 -0.261564023 -0.19306384
a2    1.000000000  0.009759915 -0.17628704 -0.18675894 -0.133518480  0.03620621
a3    0.009759915  1.000000000  0.03336910 -0.14161095 -0.196900051  0.03906025
a4   -0.176287038  0.033369102  1.00000000 -0.10131827 -0.084884259  0.07114638
a5   -0.186758940 -0.141610948 -0.10131827  1.00000000  0.388608955 -0.05149346
a6   -0.133518480 -0.196900051 -0.08488426  0.38860896  1.000000000 -0.03033428
a7    0.036206205  0.039060248  0.07114638 -0.05149346 -0.030334277  1.00000000
```

使用 cor()函数的功能是产生变量之间的相关值矩阵，如上图所示。设定参数 use="complete.obs"时，R 在计算相关值时忽略含有 NA 的记录。相关值在 1 周围表示相应的两个变量之间有强正线性相关关系。然后其他 R 函数可以得到变量间线性相关的近似函数形式，它可以让我们通过一个变量的值计算出另一个变量值。通过函数 symnum()来改善结果的输出形式：

```
> symnum(cor(algae[,4:18],use="complete.obs"))
     mP mO Cl NO NH o P Ch a1 a2 a3 a4 a5 a6 a7
mxPH 1
mnO2    1
Cl         1
NO3           1
NH4           , 1
oPO4    .  .        1
PO4     .  .        * 1
Chla .                  1
a1         .        .  .    1
a2   .                    .     1
a3                                1
a4      .           .  .             1
a5                                      1
a6            .  .                      .  1
a7                                           1
attr(,"legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
> |
```

变量 NH4 和 NO3 之间，变量 PO4 和 oPO4 之间具有相关性。样本 62 和样本 199 有太多的变量含有缺失值，若剔除它们，样本中的变量 NH4 和 NO3 就没有缺失值了。至于变量 PO4 和 oPO4，它们之间的相关性可以帮助填补这两个变量的缺失值。为此需要找到这两个变量之间的线性相关关系：

data(algae)

algae<-algae[-manyNAs(algae),]

lm(PO4 ~ oPO4,data=algae)　　# lm 用来获取线性模型，

```
> lm(PO4 ~ oPO4,data=algae)

Call:
lm(formula = PO4 ~ oPO4, data = algae)

Coefficients:
(Intercept)         oPO4
     42.897        1.293

> |
```

通过 PO4=42.897+1.293*oPO4 计算这些变量的缺失值。

```
    season   size  speed  mxPH  mnO2      Cl    NO3       NH4    oPO4     PO4    Chla   a1   a2   a3
28  autumn  small   high 6.800 11.10   9.000  0.630    20.000   4.000      NA   2.700 30.3
```

在剔除 62 和 199 后，28 在 PO4 上有缺失值，简单使用上面的线性关系计算缺失值的填补值：algae[28,"PO4"]<-42.897+1.293*algae[28,"PO4"]

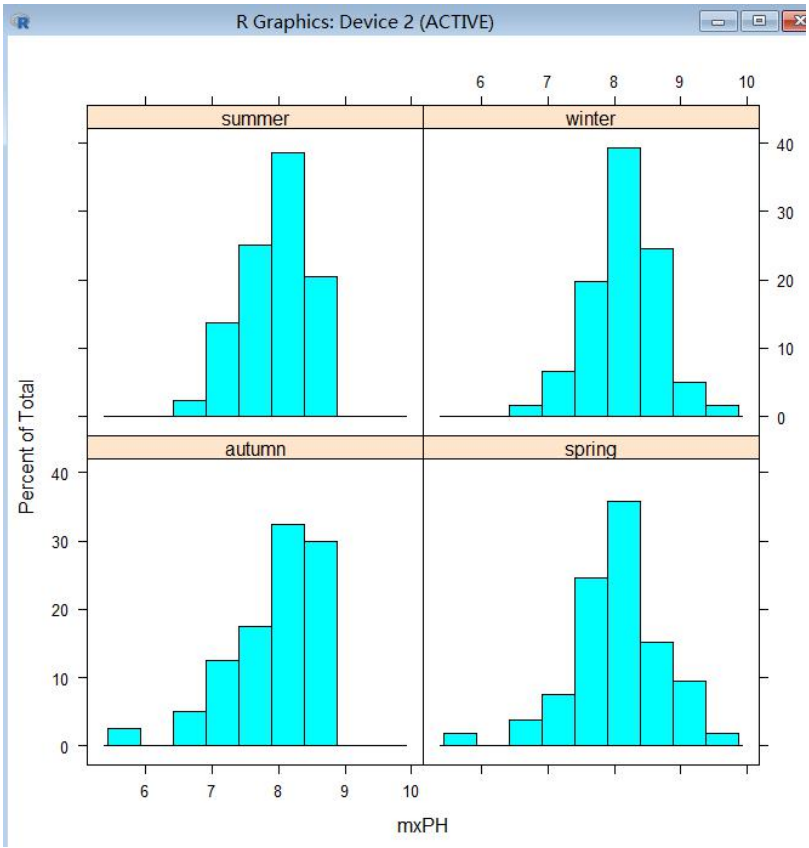
```
28  autumn  small   high 6.800 11.10   9.000  0.630    20.000   4.000  48.069   2.700 30.3
```

```
> data(algae)
> algae <- algae[-manyNAs(algae),]
> fillPO4 <- function(oP){    # 这个函数根据得到的线性关系计算变量 PO4 的值
+     if(is.na(oP))
+             return(NA)
+     else return(42.897+1.293*oP)
+ }
> algae[is.na(algae$PO4), "PO4"] <- sapply(algae[is.na(algae$PO4),
+     "oPO4"],fillPO4)          # 将这个函数应用到变量 PO4 有缺失值的所有样本，
```

结果将是填补变量 PO4 缺失值的向量。

```
> histogram(~mxPH / season, data = algae)
```

在变量 season 条件下的变量 mxPH 直方图：



(4)通过数据对象之间的相似性来填补缺失值

```
library(DMwR)
data(algae)     # 重新载入数据
```

首先显示含有缺失值的记录： algae[!complete.cases(algae),]

```
> algae[!complete.cases(algae),]
    season   size  speed mxPH mnO2     Cl   NO3 NH4    oPO4     PO4  Chla   a1   a2  a3   a4  a5  a6  a7
28  autumn  small   high 6.80 11.1 9.000 0.630  20   4.000      NA  2.70 30.3  1.9 0.0  0.0 2.1 1.4 2.1
38  spring  small   high 8.00   NA 1.450 0.810  10   2.500   3.000  0.30 75.8  0.0 0.0  0.0 0.0 0.0 0.0
48  winter  small    low   NA 12.6 9.000 0.230  10   5.000   6.000  1.10 35.5  0.0 0.0  0.0 0.0 0.0 0.0
55  winter  small   high 6.60 10.8    NA 3.245  10   1.000   6.500    NA 24.3  0.0 0.0  0.0 0.0 0.0 0.0
56  spring  small medium 5.60 11.8    NA 2.220   5   1.000   1.000    NA 82.7  0.0 0.0  0.0 0.0 0.0 0.0
57  autumn  small medium 5.70 10.8    NA 2.550  10   1.000   4.000    NA 16.8  4.6 3.9 11.5 0.0 0.0 0.0
58  spring  small   high 6.60  9.5    NA 1.320  20   1.000   6.000    NA 46.8  0.0 0.0 28.8 0.0 0.0 0.0
59  summer  small   high 6.60 10.8    NA 2.640  10   2.000  11.000    NA 46.9  0.0 0.0 13.4 0.0 0.0 0.0
60  autumn  small medium 6.60 11.3    NA 4.170  10   1.000   6.000    NA 47.1  0.0 0.0  0.0 0.0 1.2 0.0
61  spring  small medium 6.50 10.4    NA 5.970  10   2.000  14.000    NA 66.9  0.0 0.0  0.0 0.0 0.0 0.0
62  summer  small medium 6.40   NA    NA    NA  NA      NA  14.000    NA 19.4  0.0 0.0  2.0 0.0 3.9 1.7
63  autumn  small   high 7.83 11.7 4.083 1.328  18   3.333   6.667    NA 14.4  0.0 0.0  0.0 0.0 0.0 0.0
116 winter medium   high 9.70 10.8 0.222 0.406  10  22.444  10.111    NA 41.0  1.5 0.0  0.0 0.0 0.0 0.0
161 spring  large    low 9.00  5.8    NA 0.900 142 102.000 186.000 68.05  1.7 20.6 1.5  2.2 0.0 0.0 0.0
184 winter  large   high 8.00 10.9 9.055 0.825  40  21.083  56.091    NA 16.8 19.6 4.0  0.0 0.0 0.0 0.0
199 winter  large medium 8.00  7.6    NA    NA  NA      NA      NA    NA  0.0 12.5 3.7  1.0 0.0 0.0 4.9
```

> algae<-algae[-manyNAs(algae),] # 填补除去那两个含有太多 NA 值的样本外的其他缺失数据。

```
> data(algae)
> algae <- algae[-manyNAs(algae),]
> algae[!complete.cases(algae),]
    season   size  speed mxPH mnO2     Cl   NO3 NH4    oPO4     PO4  Chla   a1   a2  a3   a4  a5  a6  a7
28  autumn  small   high 6.80 11.1 9.000 0.630  20   4.000      NA  2.70 30.3  1.9 0.0  0.0 2.1 1.4 2.1
38  spring  small   high 8.00   NA 1.450 0.810  10   2.500   3.000  0.30 75.8  0.0 0.0  0.0 0.0 0.0 0.0
48  winter  small    low   NA 12.6 9.000 0.230  10   5.000   6.000  1.10 35.5  0.0 0.0  0.0 0.0 0.0 0.0
55  winter  small   high 6.60 10.8    NA 3.245  10   1.000   6.500    NA 24.3  0.0 0.0  0.0 0.0 0.0 0.0
56  spring  small medium 5.60 11.8    NA 2.220   5   1.000   1.000    NA 82.7  0.0 0.0  0.0 0.0 0.0 0.0
57  autumn  small medium 5.70 10.8    NA 2.550  10   1.000   4.000    NA 16.8  4.6 3.9 11.5 0.0 0.0 0.0
58  spring  small   high 6.60  9.5    NA 1.320  20   1.000   6.000    NA 46.8  0.0 0.0 28.8 0.0 0.0 0.0
59  summer  small   high 6.60 10.8    NA 2.640  10   2.000  11.000    NA 46.9  0.0 0.0 13.4 0.0 0.0 0.0
60  autumn  small medium 6.60 11.3    NA 4.170  10   1.000   6.000    NA 47.1  0.0 0.0  0.0 0.0 1.2 0.0
61  spring  small medium 6.50 10.4    NA 5.970  10   2.000  14.000    NA 66.9  0.0 0.0  0.0 0.0 0.0 0.0
63  autumn  small   high 7.83 11.7 4.083 1.328  18   3.333   6.667    NA 14.4  0.0 0.0  0.0 0.0 0.0 0.0
116 winter medium   high 9.70 10.8 0.222 0.406  10  22.444  10.111    NA 41.0  1.5 0.0  0.0 0.0 0.0 0.0
161 spring  large    low 9.00  5.8    NA 0.900 142 102.000 186.000 68.05  1.7 20.6 1.5  2.2 0.0 0.0 0.0
184 winter  large   high 8.00 10.9 9.055 0.825  40  21.083  56.091    NA 16.8 19.6 4.0  0.0 0.0 0.0 0.0
```

假设两个水样是相似的，其中一个水样在某些变量上有缺失值，那么该缺失值很肯能与另外一个水样的值是相似的。

用欧式距离度量相似性，使用这种度量来寻找与任何含有缺失值的数据对象最相似的 10 个水样，并用它们来填补缺失值。

可以简单的计算这 10 个最相近的数据对象的中位数并用这个中位数来填补缺失值。

还可以采用这些最相似数据的加权平均值。权重的大小随着距待填补缺失值的数据对象的距离增大而减小，

以上描述的方法可通过 knnImputation()函数来实现：

```
algae<-knnImputation(algae,k=10)
```

若用中位数来填补缺失值：

algae<-knnImputation(algae,k=10,meth="median")

```
> algae<-knnImputation(algae,k=10)
> algae[!complete.cases(algae),]
 [1] season size   speed  mxPH   mnO2   Cl     NO3    NH4    oPO4   PO4    Chla   a1     a2     a3     a4     a5     a6
[18] a7
<0 行> (或0-长度的row.names)
```

通过这些操作数据集中不再含有 NA 值(缺失值)。

4、实验环境：

使用 R 软件的 Windows 版本，运行下载文件 R-3.3.0-win.exe。

安装实验中需要用到的包：

installed.packages('DMwR')