**数据介绍：**

有200个水样，每条记录是同一条河流在该年的同一个季节的三个月内收集的水样的平均值。

每条记录由11个变量构成，3个是标称变量，分别描述水样收集的季节，河流大小和河水速度，剩下的8个变量是水样的化学参数：

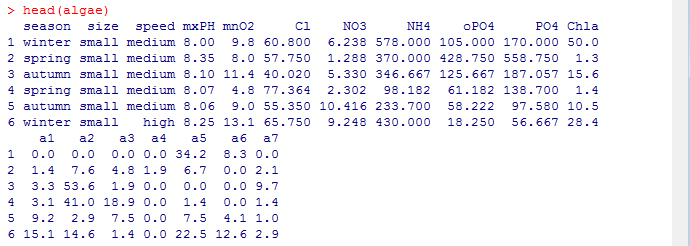
* 最大pH值(mxPH)
* 最小含氧量(mnO2)
* 平均氯化物含量(Cl)
* 平均硝酸盐含量(NO3)
* 平均氨含量(NH4)
* 平均正磷酸盐含量(oPO4)
* 平均磷酸盐含量(PO4)
* 平均叶绿素含量(Chla)

a1-a7为7种不同有害藻类在相应水样中的频率数目。

## 读取数据

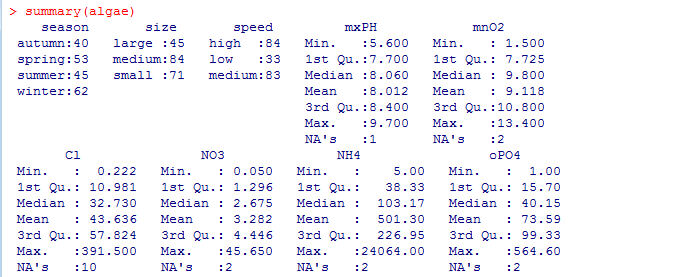
algae <- read.table(file = 'Analysis.txt', col.names = c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'), na.strings = c('XXXXXXX'))

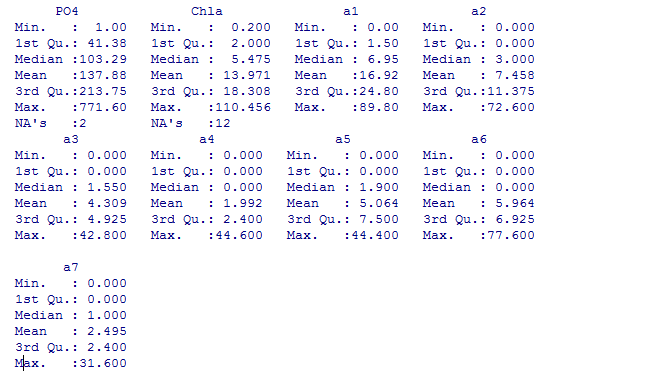
确定读入成功：



## 数据摘要和可视化

### 2.1 数据摘要



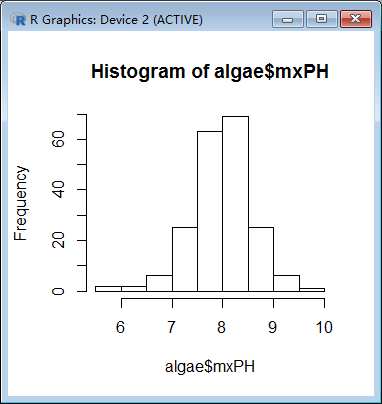


这个简单函数给出了数据的的统计特征概括，对于名义变量，他给出了每个取值的变量的频数。例如，从结果中可知冬季采集的水样比其他季节更多，河流size为最大的有45个，河流的流速low比较少。对于数值型变量，R为我们提供了四分之一位数，中位数，均值，四分之三位数，极值等一系列信息。这些统计信息提供了变量值分布的初步信息，在变量有缺失值的情况下，字符串NA后面的数值即为缺失值的个数，通过中位数，均值，四分位数的信息，我们可以了解数据分布的偏度和分散情况。且这些信息大多数都可以通过图形来表达出来。

### 2.2 数据可视化

（1）绘制直方图：mxPH

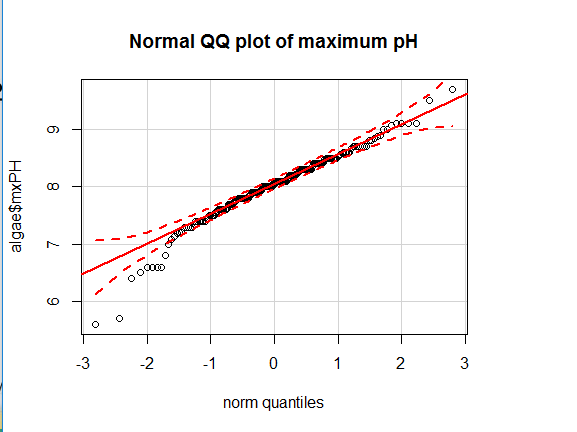
hist(algae$mxPH)



（2）绘制Q-Q图：mxPH

> par(mfrow=c(1,2))

> qq.plot(algae$mxPH,main="Normal QQ plot of maximum pH")



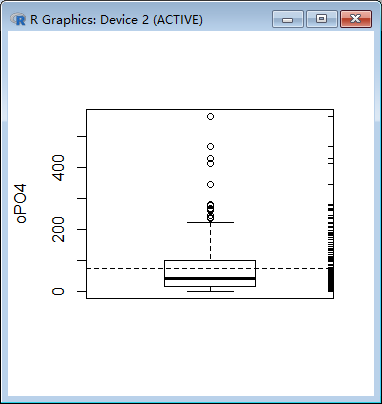
绘制了变量值与正态分布的理论分位数的散点图，同时给出正态分布的95%的置信区间的带状图，从上图知，变量有几个小的值明显在95%置信区间之外，它们不服从正态分布。

（3）绘制盒图：oPO4

> boxplot(algae$oPO4,ylab='oPO4')

> rug(algae$oPO4,side=4)

> abline(h=mean(algae$oPO4,na.rm=T),lty=2)



盒图上方小横线上面的小圆圈表示与其他值比较特别大的值，通常认为是离群值

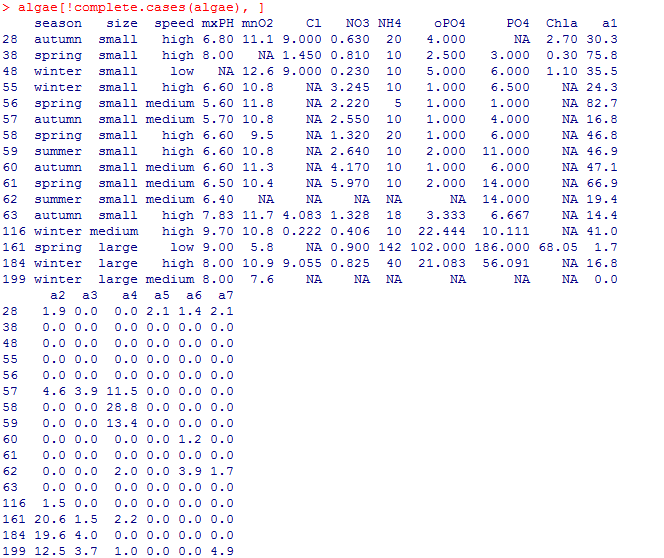
## 数据缺失的处理

### 3.1 剔除缺失值

剔除含有缺失数据的记录非常容易实现，尤其是当这些记录所占的比例在可用数据集中非常小的时候，这个选择就比较合理。

1. 检查含缺失值的记录

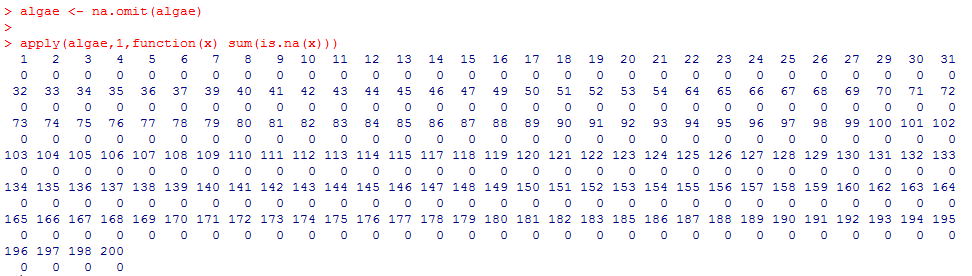
> algae[!complete.cases(algae), ]



2. 剔除缺失值

> algae <- na.omit(algae)

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



3.保存预处理后的数据集

> write.table(algae,file = 'Analysis\_1.txt',quote=FALSE,sep='\t',row.names =FALSE, col.names = FALSE)

### 3.2 用最高频率值来填补缺失值

填补含有缺失值记录的另一种方法是尝试找到这些缺失值最可能的值。

1. 使用centralImputation()自动用数据中心趋势填补数据值所有缺失值。

> library(DMwR)

> algae<-centralImputation(algae)

2. 保存预处理后的数据集

> write.table(algae,file = 'Analysis\_2.txt',quote=FALSE,sep='\t',row.names =FALSE, col.names = FALSE)

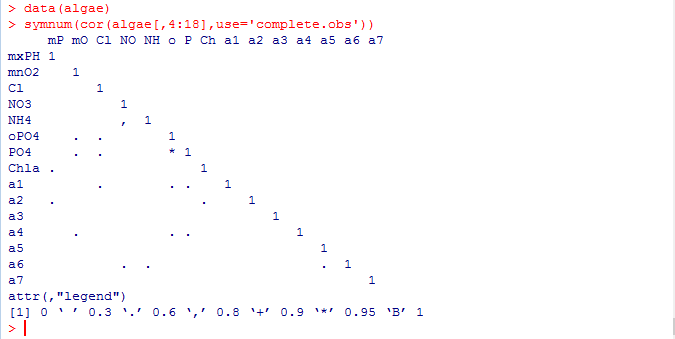
### 3.3 通过属性的相关关系来填补缺失值

另一种获取缺失值较少偏差估计值的方法是探寻变量之间的相关关系。比如，通过变量值之间的相关关系，能够发现某变量与mxPH高度相关。

1. 获取变量之间的相关矩阵

> data(algae)

> symnum(cor(algae[,4:18],use='complete.obs'))



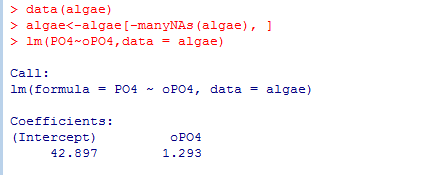
结果显示，大多数变量之间是不相关的，然而，有两个例外：变量NH4和NO3之间，变量PO4和oPO4之间。后两个变量之间的相关值很高（大于0.9）。，因此，根据它们来确定缺失数据是危险的。此外，因为样本62和样本199有太多的变量含有缺失值，所以如果剔除它们，样本中的变量NH4和NO3就没有缺失值了。

2. 寻找变量PO4和oPO4之间的线性关系

> data(algae)

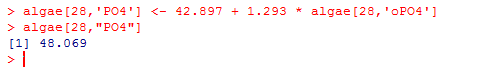
> algae<-algae[-manyNAs(algae), ]

> lm(PO4~oPO4,data = algae)



线性模型是 PO4=1.29oPO4+42.90，如果这两个变量不是同时有缺失值，那么可以用这个公式计算这些变量的缺失值。

3. 在剔除样本62和样本199后，还剩下一个样本28 在PO4上有缺失值，可以简单地使用上面的线性关系计算缺失值的填补值



通过变量之间的相关关系求出水样28，PO4的缺失值填补为48

4. 保存预处理后的数据集

> write.table(algae,file = 'Analysis\_3.txt',quote=FALSE,sep='\t',row.names =FALSE, col.names = FALSE)

### 3.4 通过数据对象之间的相似性来填补缺失值

不同于探索数据集列之间的相关性，尝试使用行之间的相似性来填充缺失值。我们可以使用这种方法来填补剔除了那两个含有太多的NA值得样本外的其它缺失数据。

如上所描述的方法假如两个水样是相似的，其中某些变量有缺失值，那么该缺失值很可能与另一个水样的值是相似的。为了使用这种直观的方法，首先定义相似性的概念。相似性经常由描述观察值的多元度量空间的变量所定义。在文献中有许多度量相似性的指标，常用的是欧式距离。这个距离可以非正式的定义为任何两个案例的观察值之差的平方和。

下面要描述的方法是使用这种度量来寻找与任何含有缺失值的案例最相似的10个水样，并用它们来填充缺失值。我们考虑两种应用这些值的方法。第一种方法简单地计算这10个最相近的案例的中位数并用这个中位数来填充这些缺失值。第二种方法是采用这些相似数据的加权均值。权重的大小随着距待填补缺失值的个案的距离为d，则它的值在加权平均中的权重为：w d =𝑒−𝑑

1. 上面的方法可以通过添加包函数knnImputation()来实现，这个函数用一个欧式距离的变种来找到距离任何个案最近的k个邻居

> algae <- read.table(file = 'Analysis.txt', 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'))

> library(DMwR)

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

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

2. 保存预处理后的数据集

> write.table(algae,file = 'Analysis\_4.txt',quote=FALSE,sep='\t',row.names =FALSE, col.names = FALSE)