数据挖掘作业1——R语言下的海藻数据分析

苏思悦 2120151031

1. **系统环境**

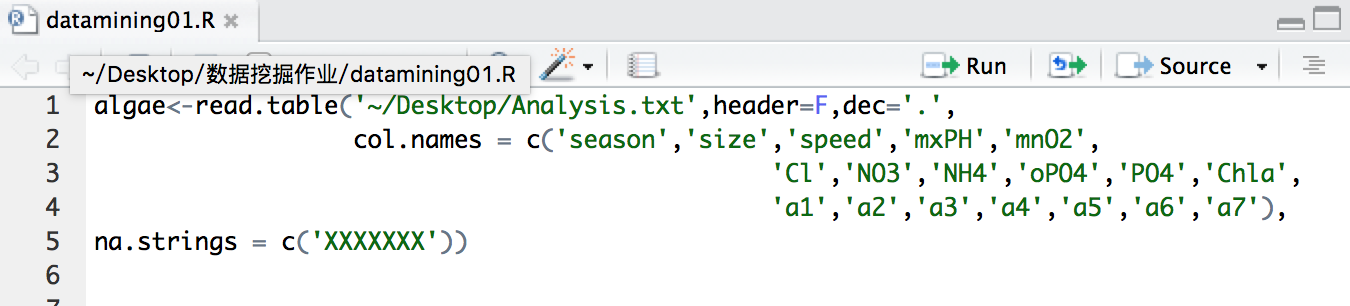
本作业是在MAC系统下，使用R语言编程

1. **2、从文件中读入数据**

从网上下载了原始的数据文件Analysis.txt，保存在桌面上，文件的路径为

~/Desktop/Analysis.txt

在R语言的编译器下通过以下代码读入数据文件：

****

1. **数据摘要**

获取数据统计特性的方法之一是获取其描述性的统计摘要

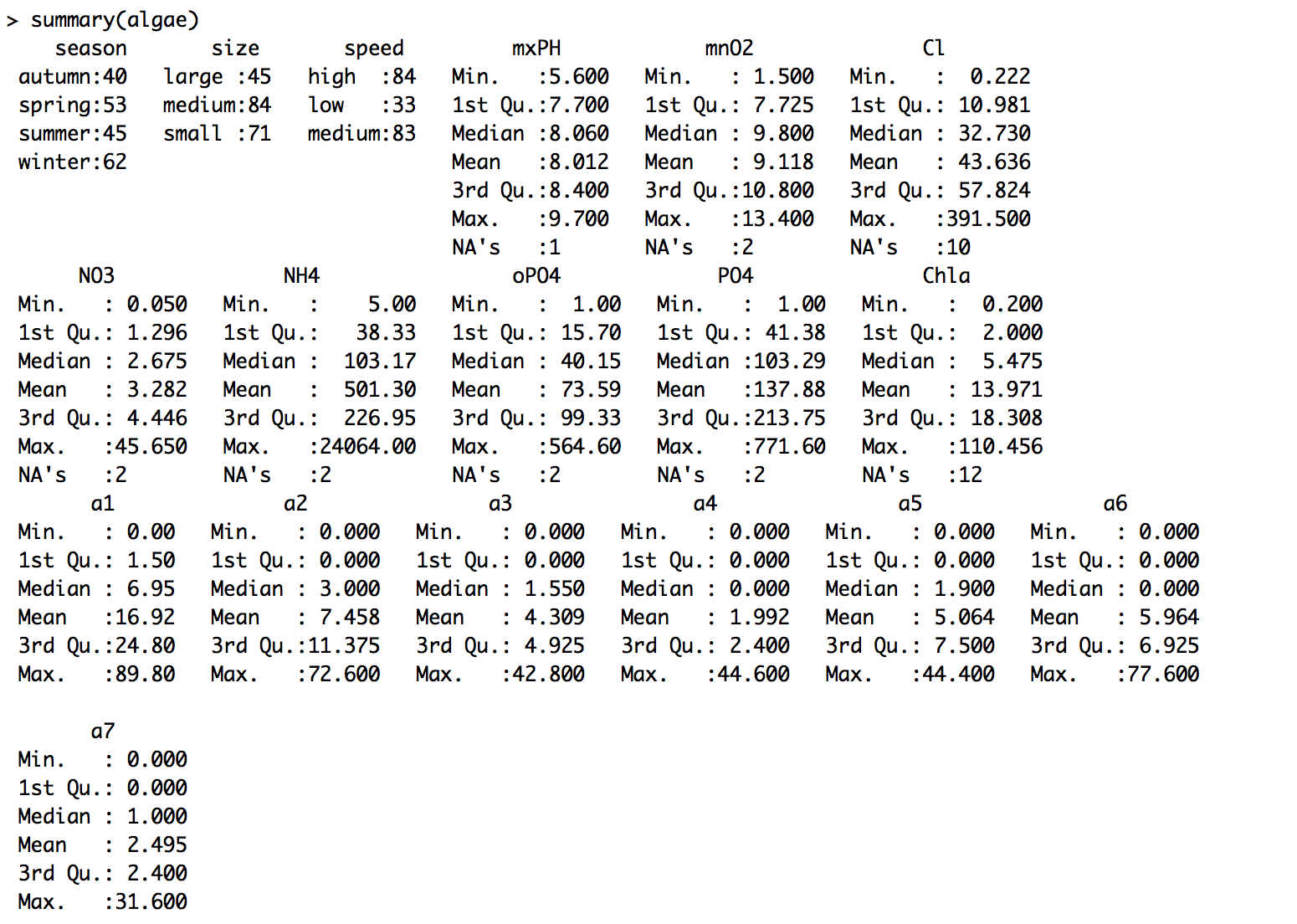
采用以下命令：summary(algae)

对于标称变量，可以得到每个可能取值的频数

对于数值变量，可以得到其均值、中位数、四分位数和极值。

其中，NA‘s表示缺失值的个数。

数据摘要如下：



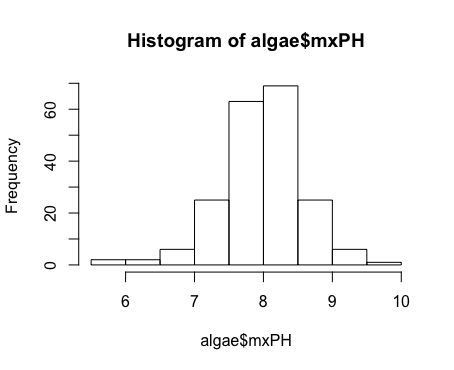
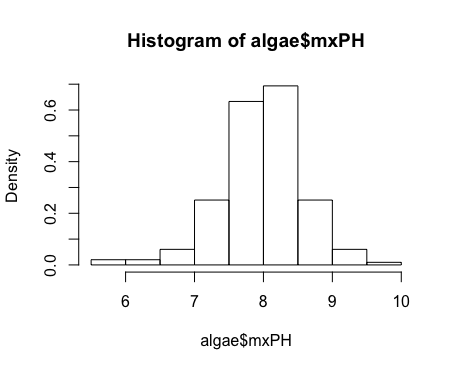
1. **数据可视化**

4.1 绘制直方图

绘制直方图的方式有两种，以变量mxPH为例：

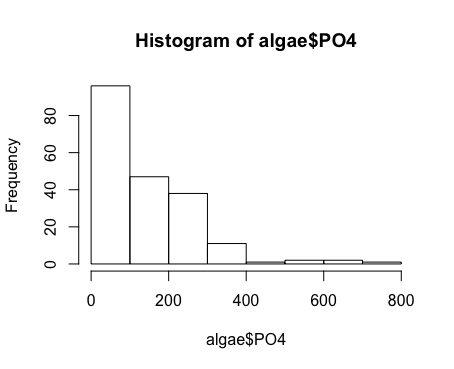
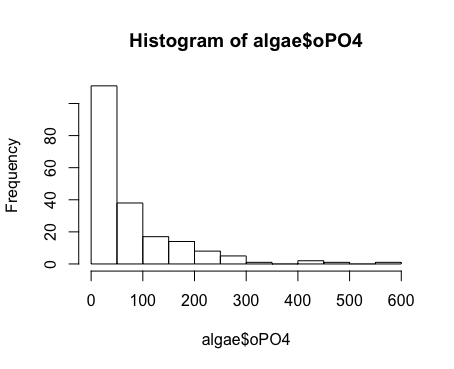
1. hist(algae$mxPH) （给出的是频数）
2. hist(algae$mxPH,prob=T) （给出的是区间的概率）

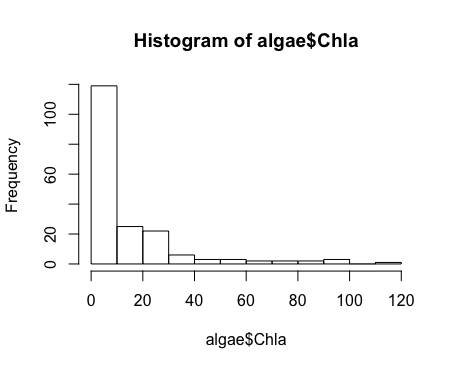
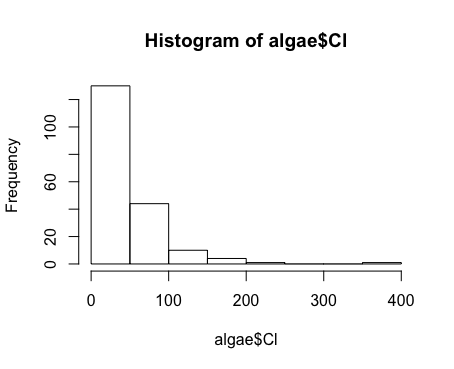
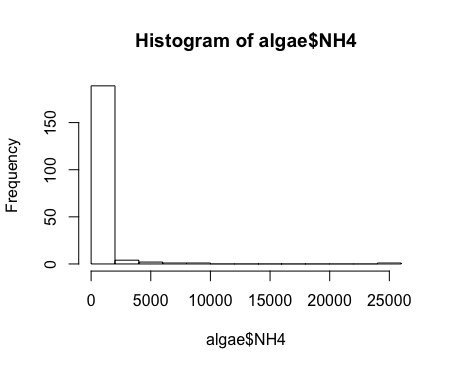
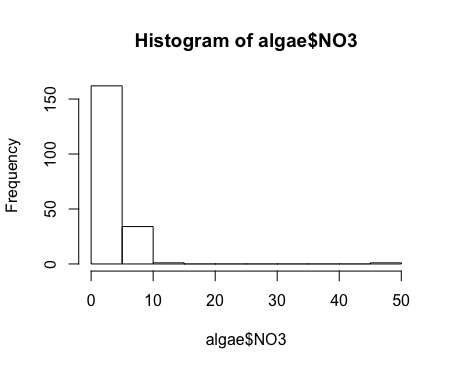
绘制的直方图：

** **

**hist(algae$mxPH) hist(algae$mxPH,prob=T)**

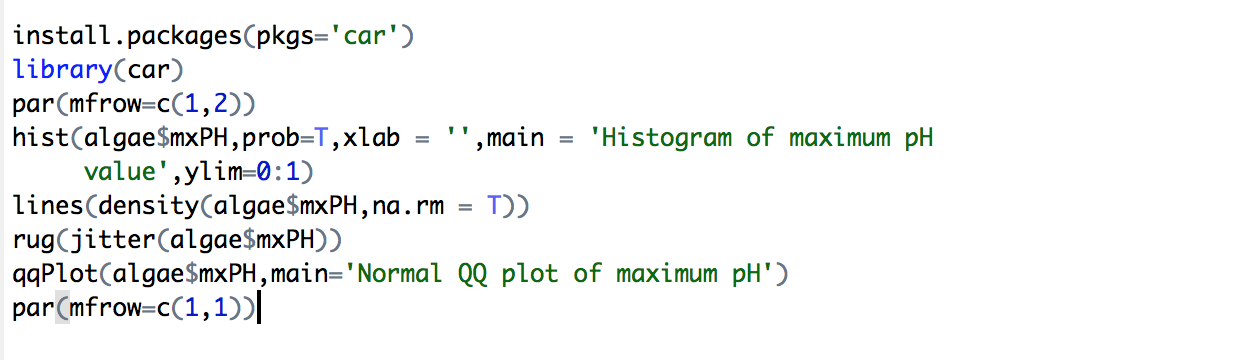
**其他变量的直方图：**

** **

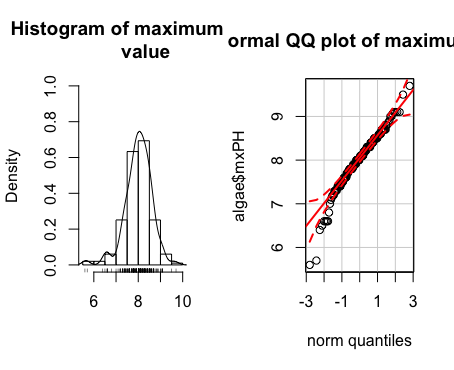
****

**4.2 用qq图检验是否为正态分布**

**以变量mxPH为例，用以下代码检验其是否为正态分布**



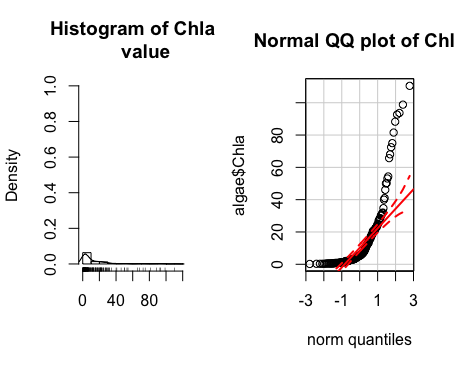
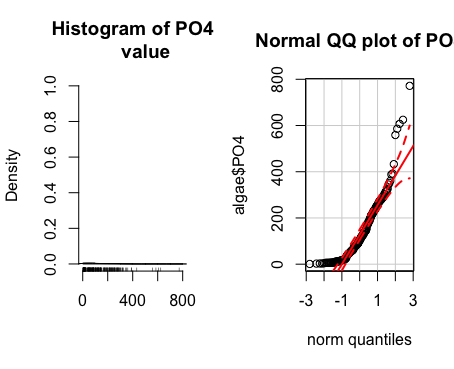
**图：**

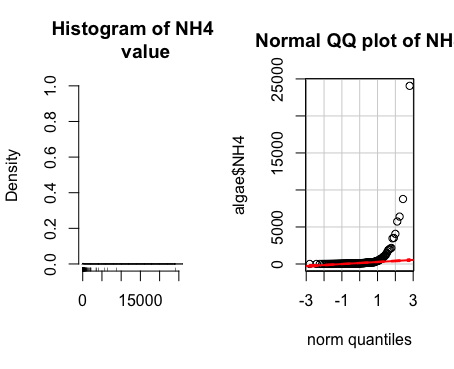
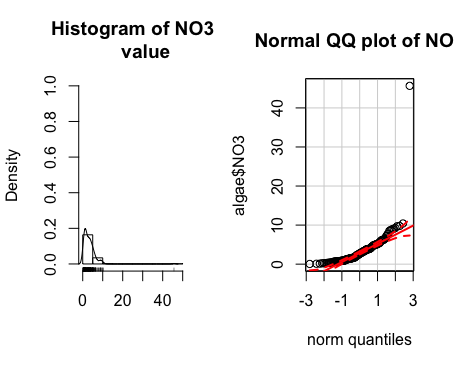
****

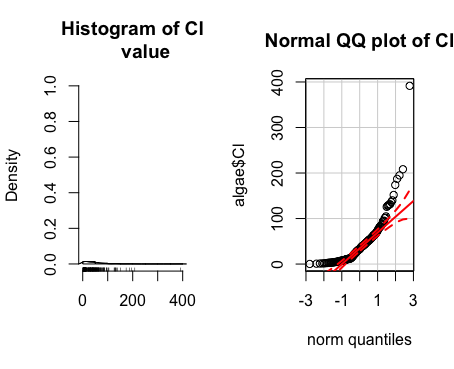
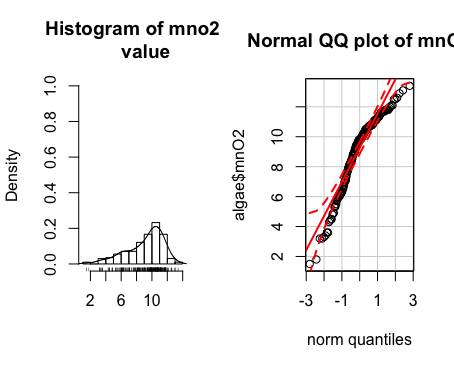
**qq图绘制变量值和正态分布的理论分位数的散点图。同时，它给出正态分布的95%置信区间的带状图。main为设置图形的标题。**

**变量mxPH的值符合正态分布**

**其他变量的正态分布图：**

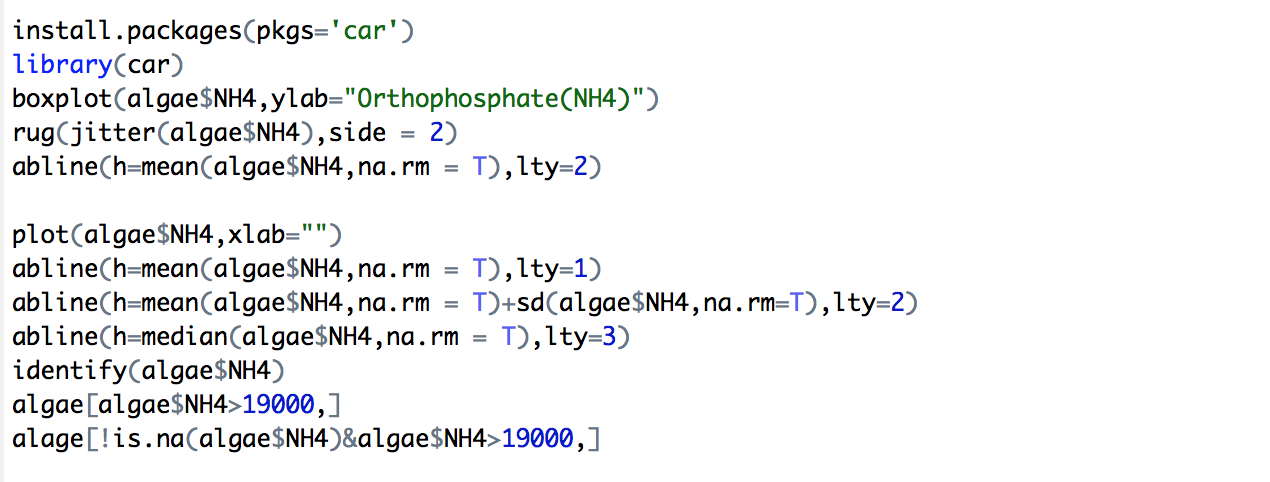
****

****

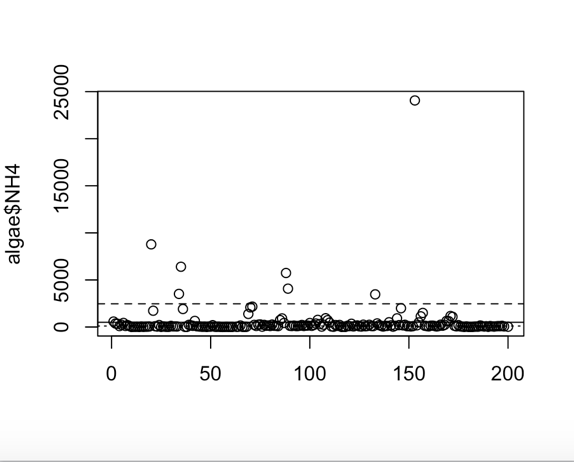
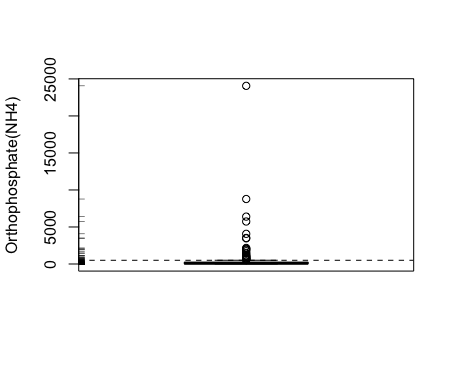
****

**4.3 盒图与离群值识别**

**以变量NH4为例，用以下代码绘制盒图并进行离群点标示**

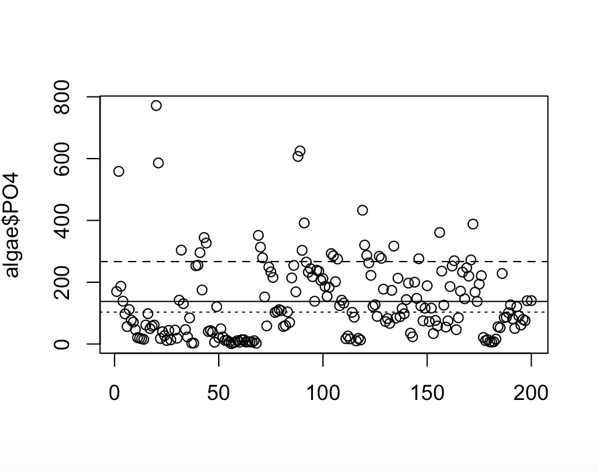
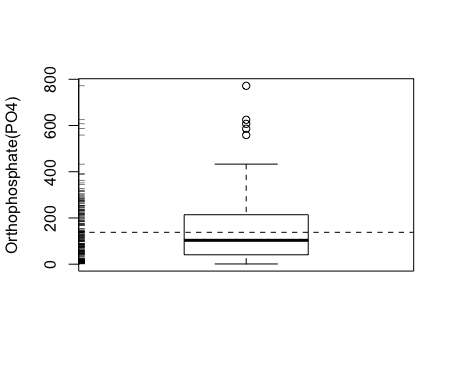
****

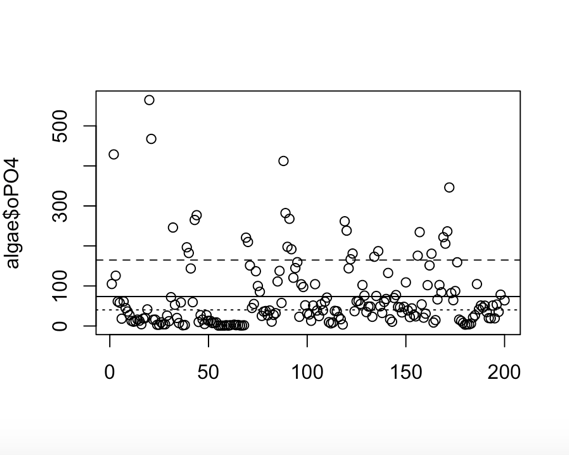
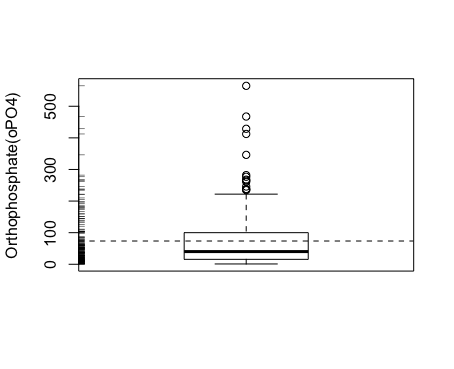
如图：

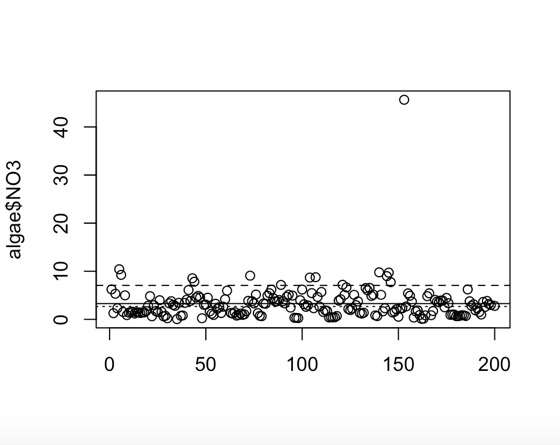
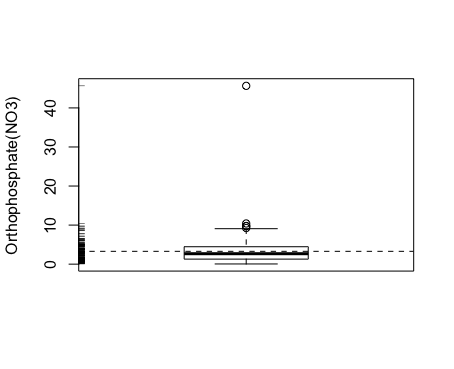
****

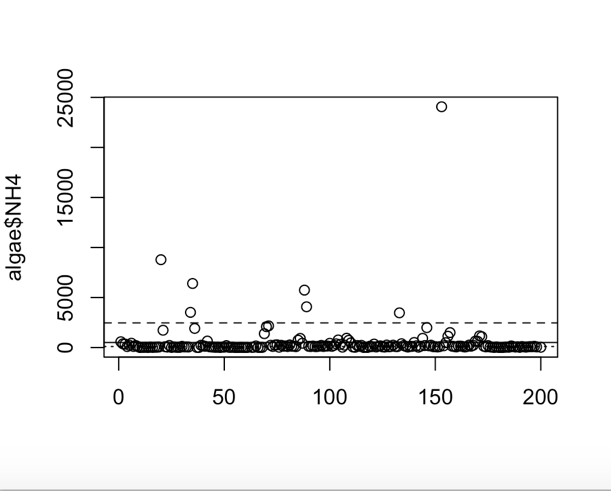
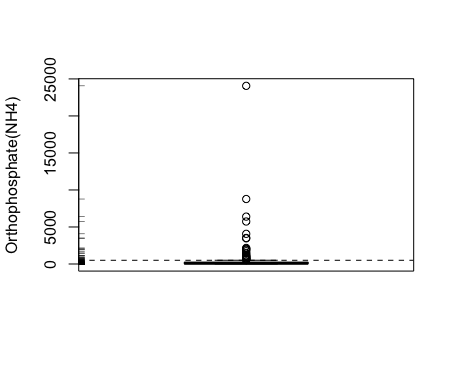
**盒图 离群点标示**

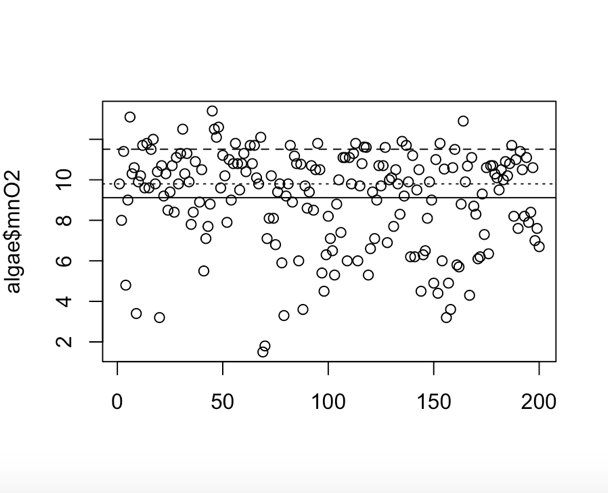
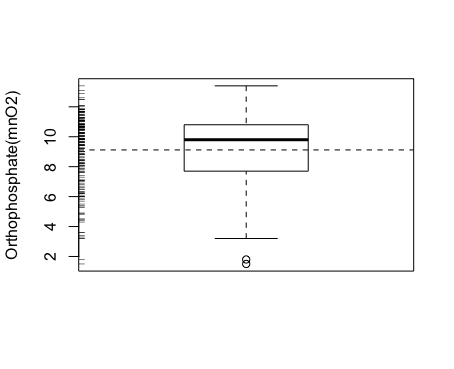
**其他变量：**

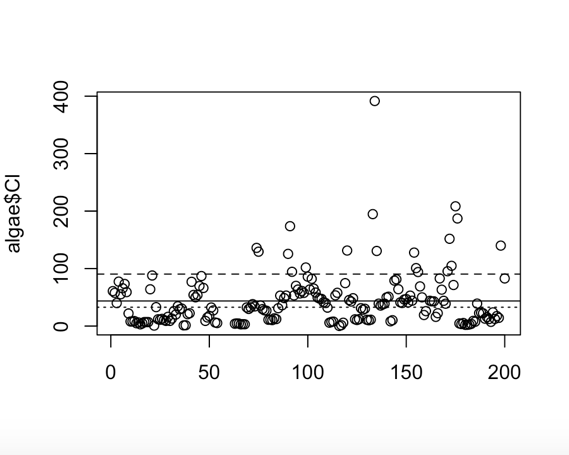
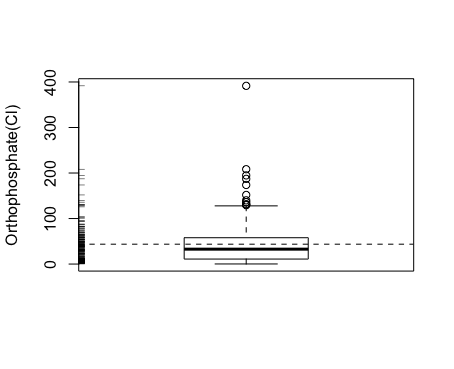
****

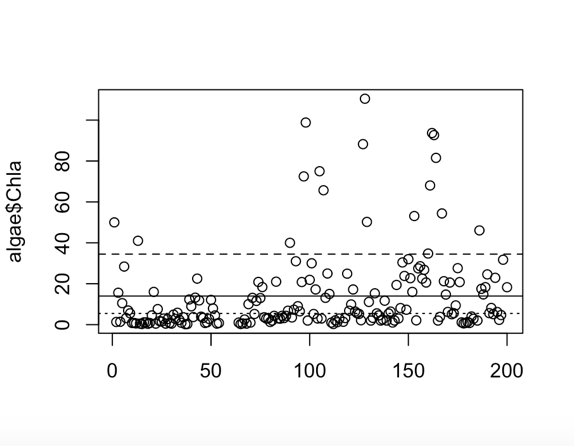
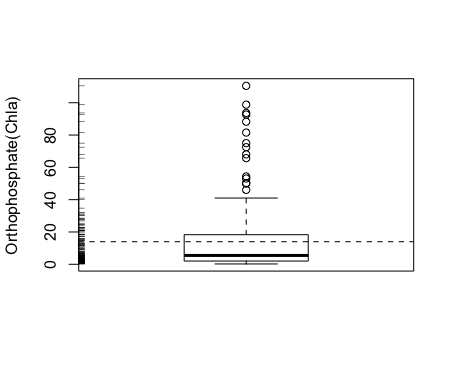
****

****

****

****

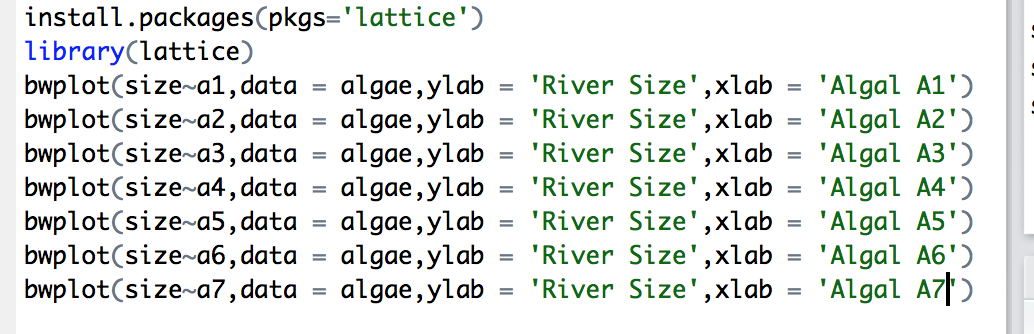
****

****

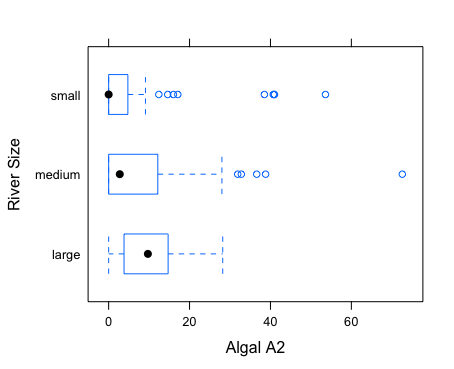
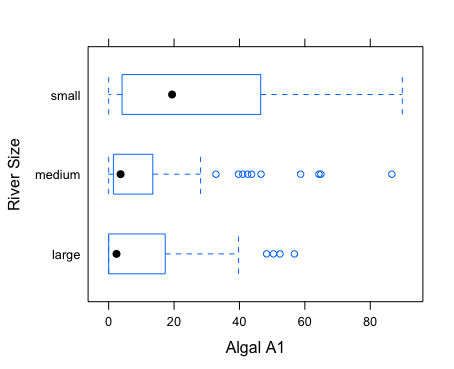
**4.4 条件盒图**

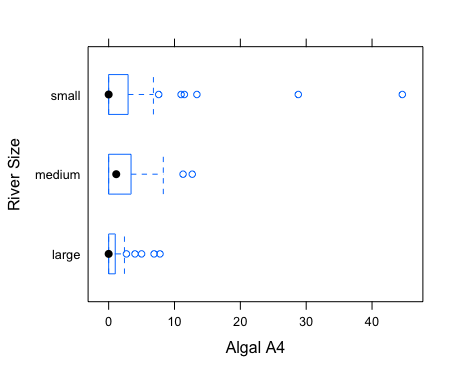
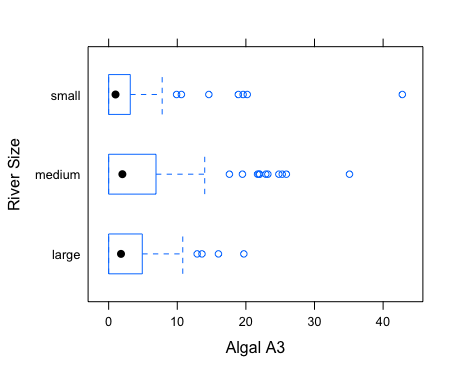
**对于7种海藻，绘制其数量与标称变量。**

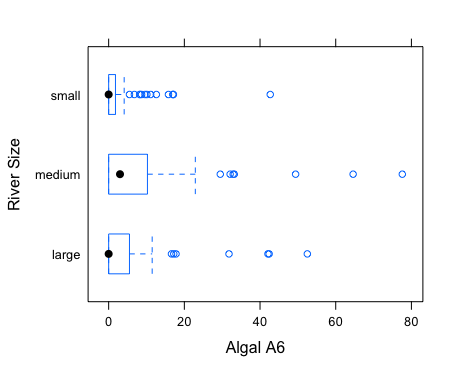
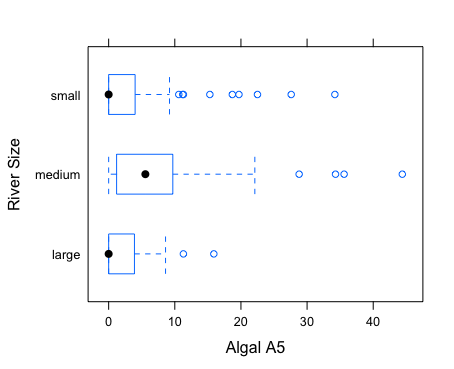
**代码如下：**

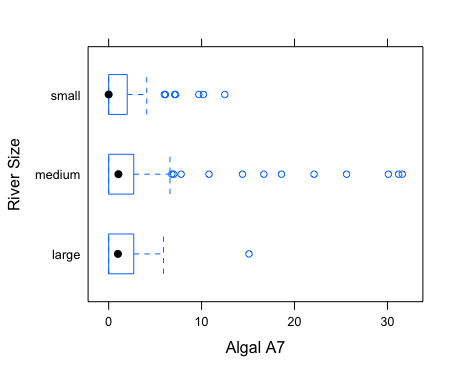
****

**以size的条件盒图为例：**

****

****

****

****

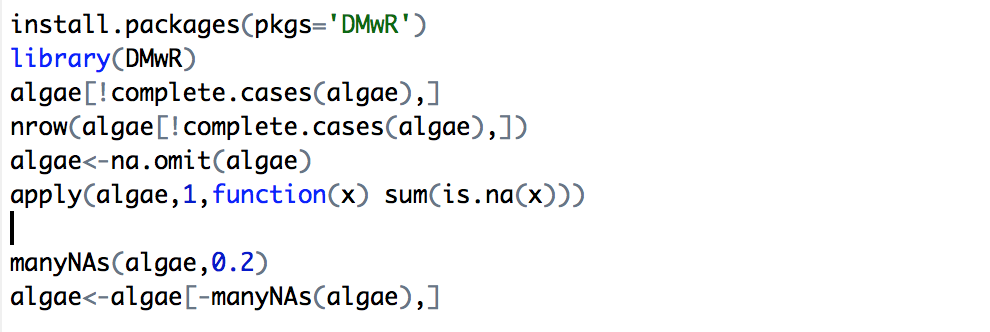
1. **数据缺失的处理**

分别使用下列四种策略对缺失值进行处理:

* 将缺失部分剔除
* 用最高频率值来填补缺失值
* 通过属性的相关关系来填补缺失值
* 通过数据对象之间的相似性来填补缺失值

**5.1**将缺失部分剔除

**代码**



algae[!complete.casea(algae),]：检查含缺失值的记录

algae<-na.omit(algae)：剔除所有含缺失值的记录

apply(algae,1,function(x) sum(is.na(x)))：找出每个记录中缺失值的个数

apply()函数里的第二个参数‘1’表示第一个参数algae中的对象的第一个维度，即行数据。第三个参数是临时函数，功能是计算对象x中NA的数量。

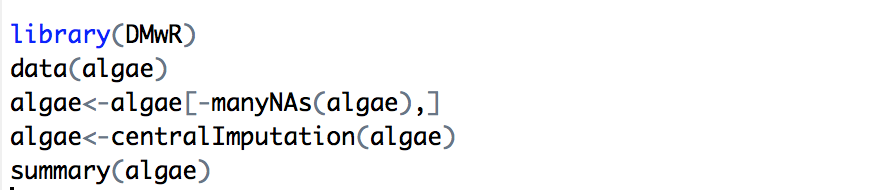
library(DMwR)

manyNAs(algae,0.2):第二个参数如不指定，默认为0.2。

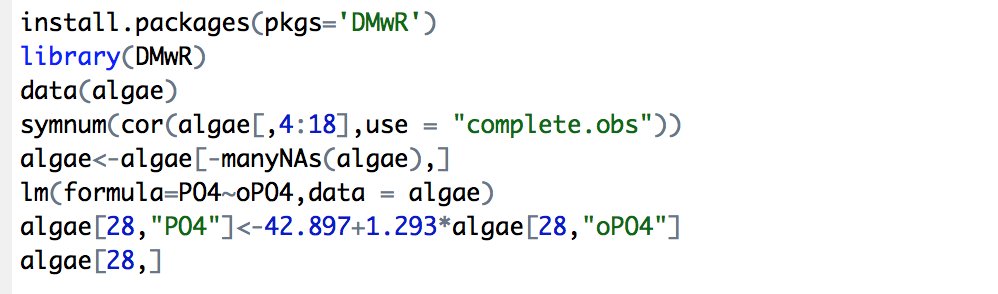
algae<-algae[-manyNAs(algae),]:剔除缺失值较多的记录。

**5.2**用最高频率值来填补缺失值

代表中心趋势的值反映了变量分布的最常见值。对于接近正态的分布来说，均值是最佳选择；对偏态分布或有离群值的分布而言，中位数通常是更好的代表数据中心趋势的指标；对于名义变量，通常采用众数。用以下函数完成填补所有缺失值：



**5.3**通过属性的相关关系来填补缺失值

****

data(algae)

symnum(cor(algae[,4:18],use = "complete.obs"))：其中，函数cor()产生相关值矩阵，use参数指计算相关值时忽略含有NA的记录。

结果显示，NH4和NO3，PO4和oPO4之间的相关性较大。

NH4和NO3：相关性为（0.6～0.8）剔除样本62和样本199（含有较多的缺失值）。

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

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

可以得到线性模型：PO4=42.897+1.293\*oPO4，由于样本62和样本199已经剔除，所以仅样本28在PO4上有缺失值，可以用上面的线性关系来填补：

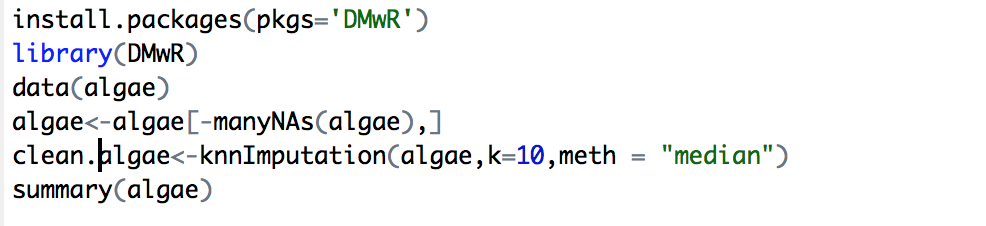
algae[28,"PO4"]<-42.897+1.293\*algae[28,"oPO4"]

algae[28,]：查看填补的记录。

**5.4**通过数据对象之间的相似性来填补缺失值

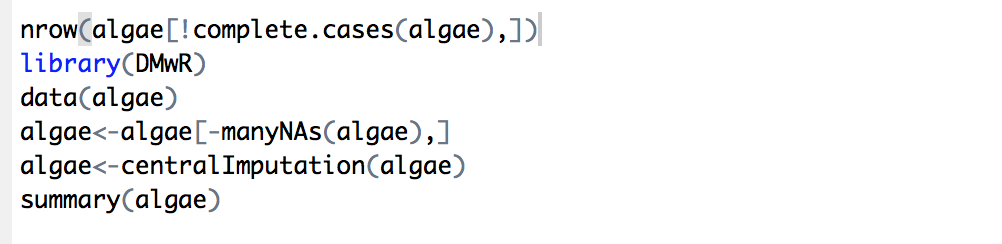
度量相似性时，最常用的指标是欧式距离。利用这种度量来寻找与任何含有缺失值案例最相似的10个水样，并用它们填补缺失值。

实验中，我们采用这些最相似数据的加权均值，用高斯核函数从距离获得权重。代码如下：

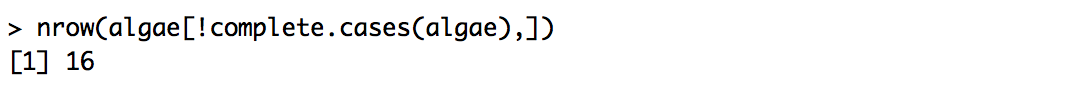


**5.5 数据预处理前后的对比**

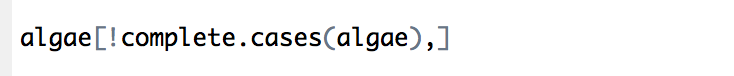
**以第二个方法为例：**

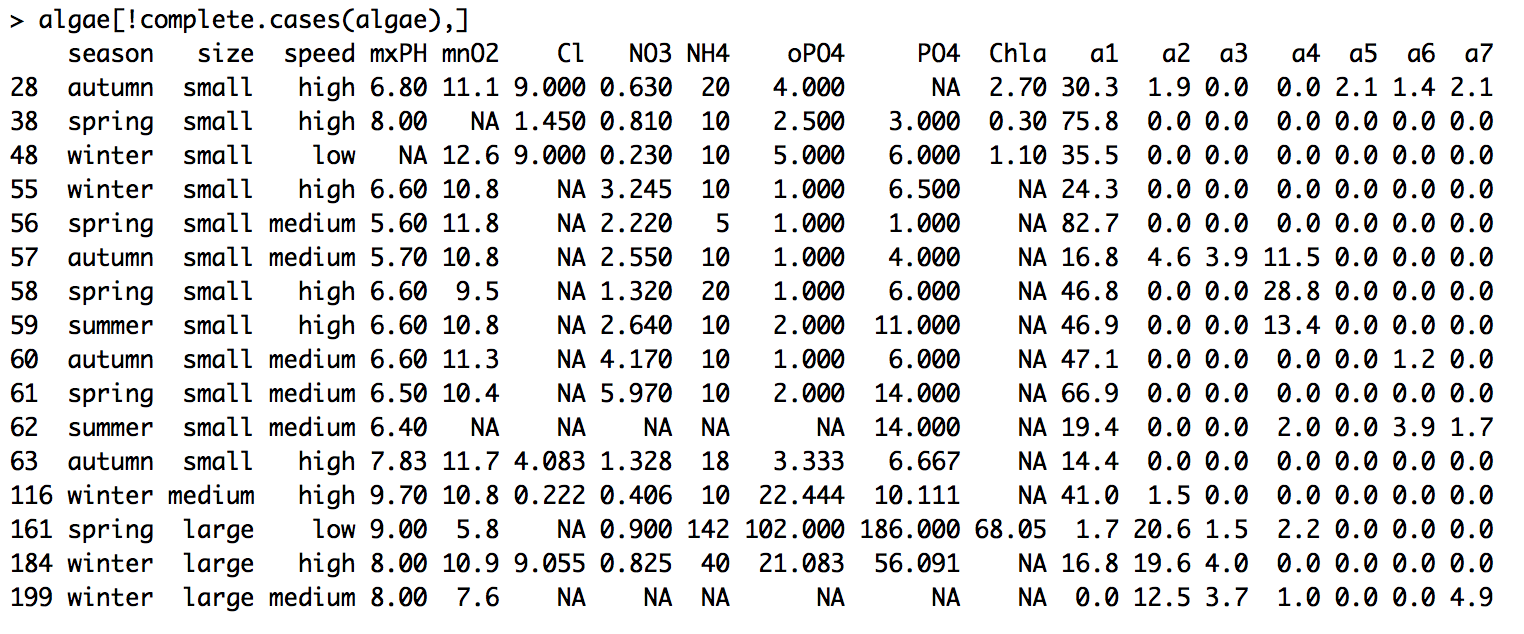
****

**用nrow(algae[!complete.cases(algae),])来检查共有多少条缺失记录：**

****

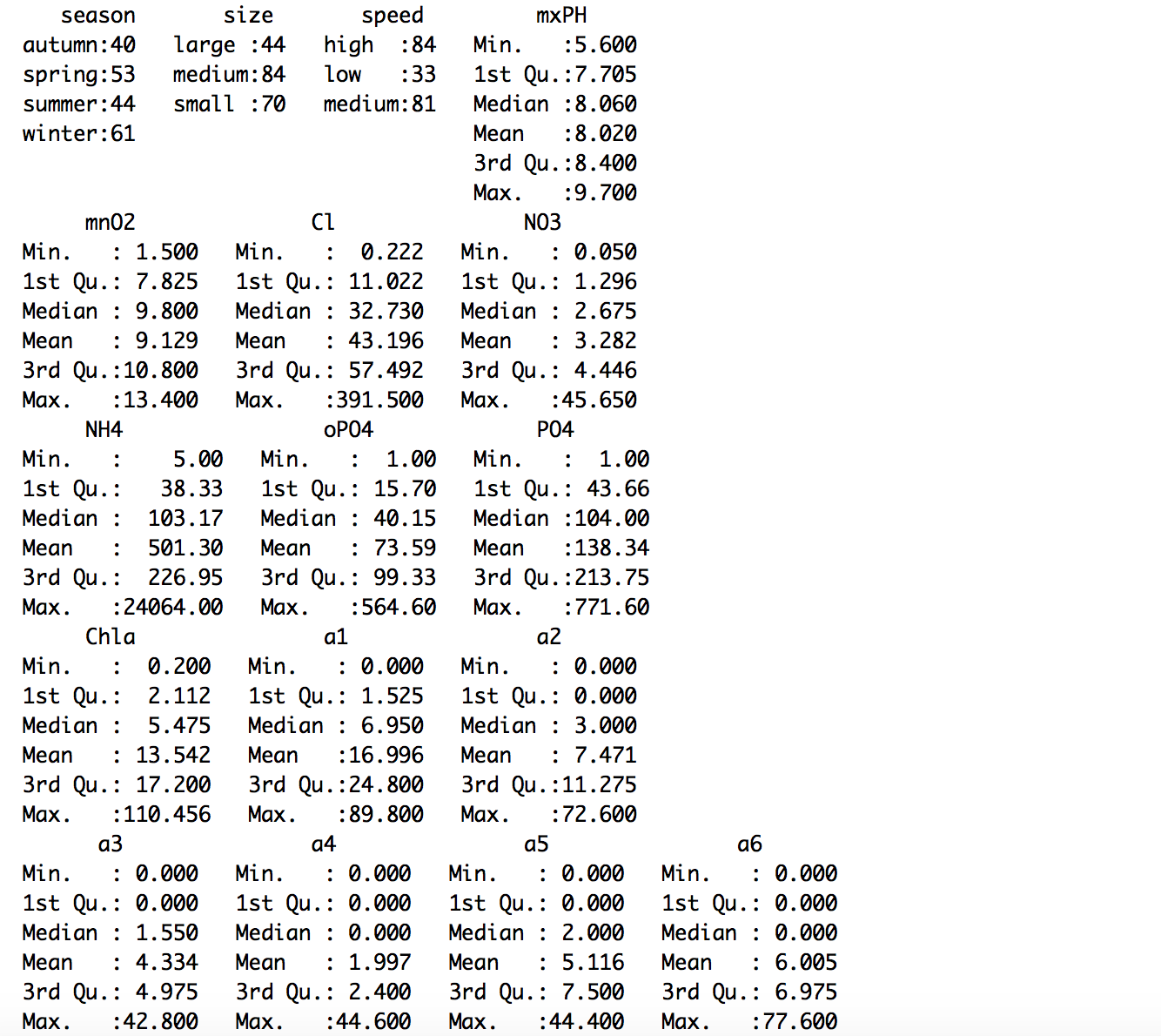
**用下面的命令去显示所有的缺失值**

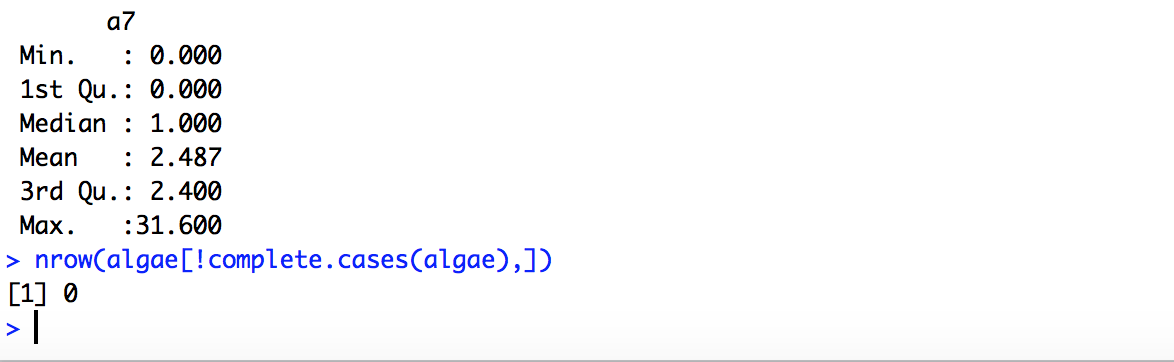
****

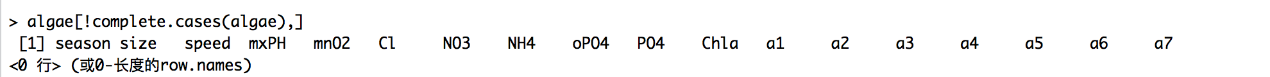
****

**程序运行后，说明有16条缺失记录。**

**在数据预处理后，数据摘要为：**

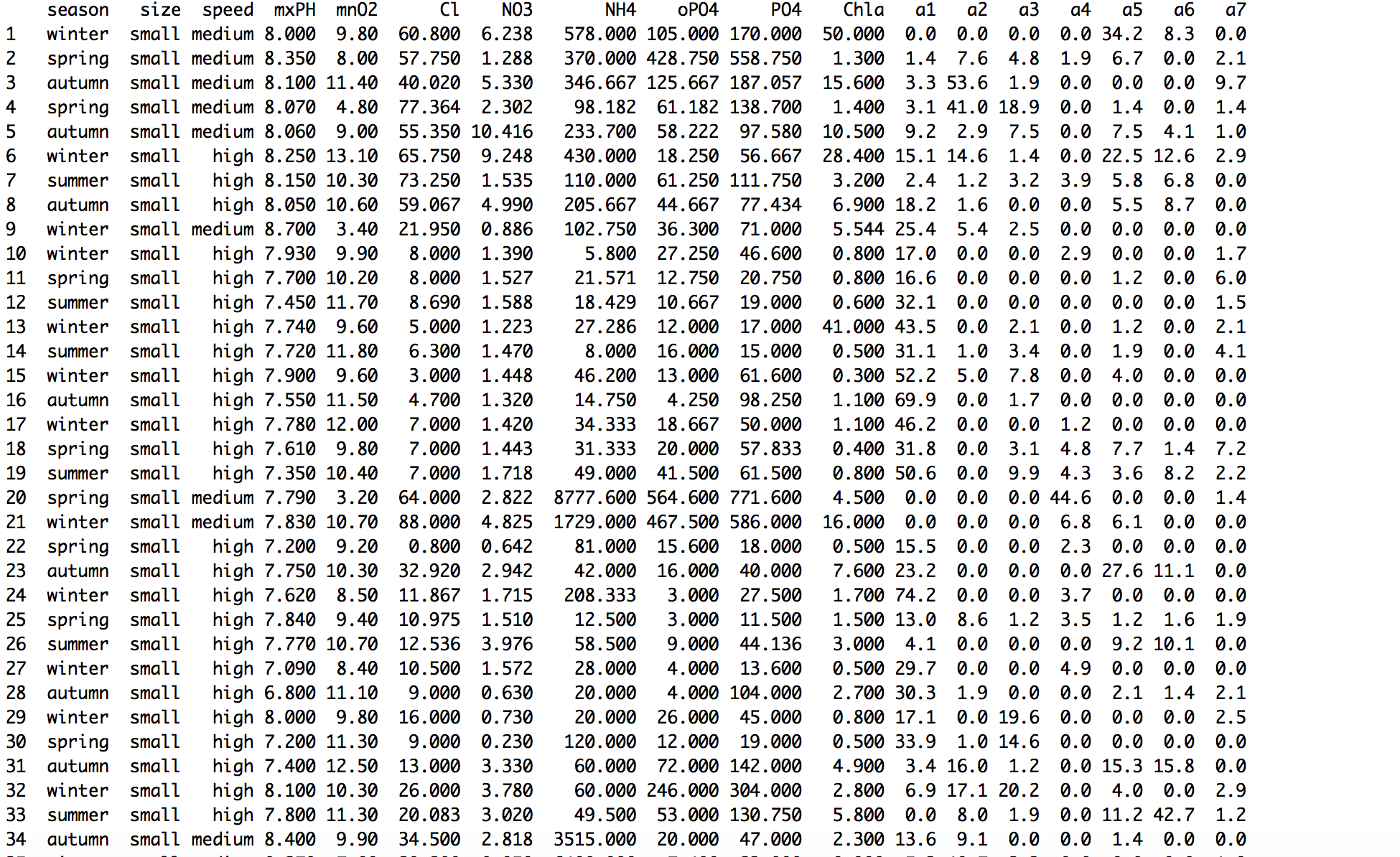
****

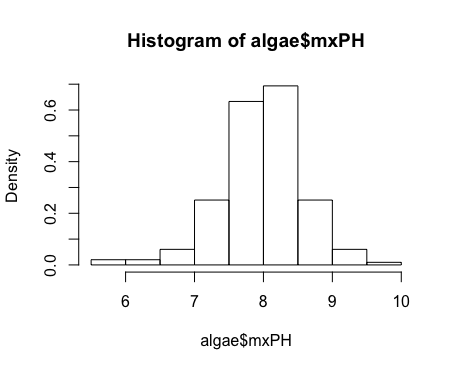
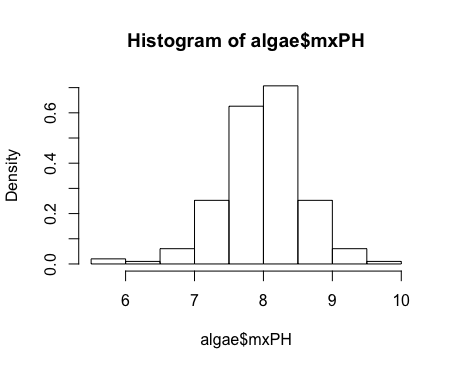
****

****

从中可以看出，调用nrow(algae[!complete.cases(algae),])后，显示缺失值为0数据集中的缺失值已经被填充。

查看处理过后的数据集，我们可以看到所有的缺失值都已经被填充，下图显示前34条数据：

****

** **

**处理前 处理后**

**5.6 数据集保存**

用下面的命令将预处理后的数据集保存，保存的目录为电脑桌面，文件名为

Results.txt。

****

1. **总结**

**通过这次的作业，我对于数据可视化和数据预处理的过程有了更好的理解，并且通过作业，对于R语言有了学习使用的机会。在以后的学习生活中我会继续使用R语言进行数据挖掘的工作。**