# 数据挖掘 海藻数据分析

## 1 实验环境搭建

本实验使用R语言实现海藻数据分析，所需实验环境和软件如下：

1. 安装R语言软件包
2. 安装car、DMwR等分析数据所需的程序包

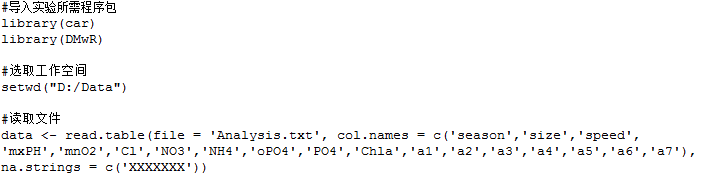
## 2 **实验过程**

## 2.1 数据可视化与摘要

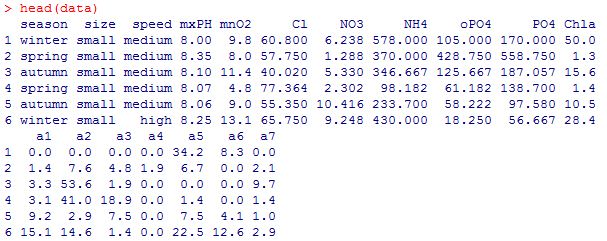
### 2.1.1 数据摘要

(1) 数据预处理

在对数据进行处理之前，需要对数据进行预处理。引入DMwR、car两个程序包，并将Analysis.txt中的数据都入到data中。



运行结果：

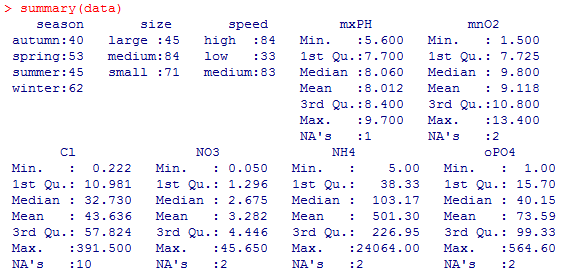


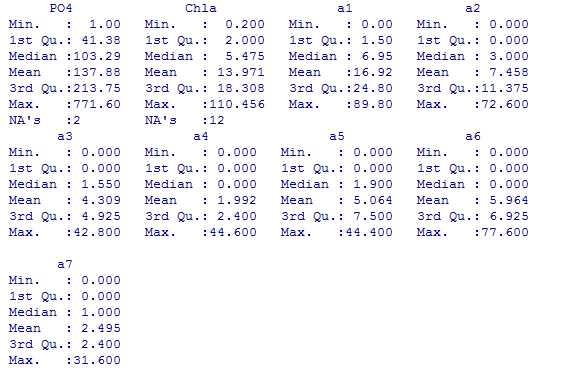
(2) 对标称属性，给出每个可能取值的频数；对于数值属性，得到最大、最小、均值、中位数、四分位及缺失值得个数。

调用R语言自带的summary函数



运行结果：





这里我们使用了R语言自带的summary()函数。可以查看对象的基本信息（min, max, mean, etc）。从得到的数据可以看出，对于标称变量（season, size, speed）给出了每个变量的频数。对于其他的数值属性，分别给出了其最小（min），中位数（Median），均值（Mean），四分位数（1st Qu, 3rd Qu），最大值（Max），在变量有缺失值的时候，NA后面的数值即为缺失值的个数。

运行结束后，将处理后的数据存储在文件attributeData.txt中。



### 2.1.2 数据可视化

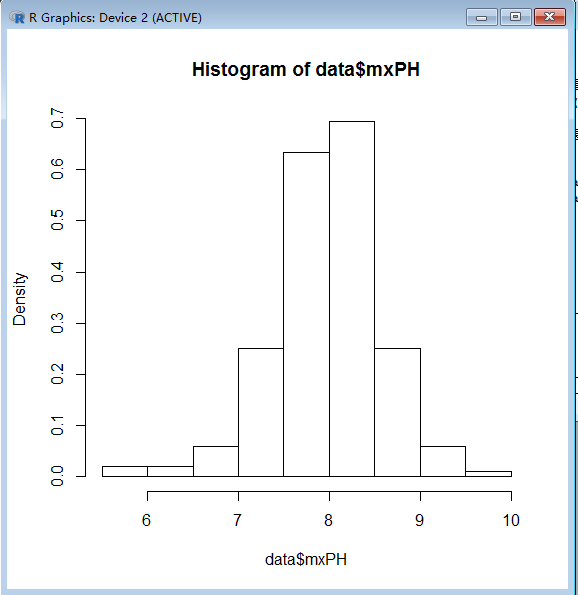
（1）针对数值属性绘制直方图

调用hist函数

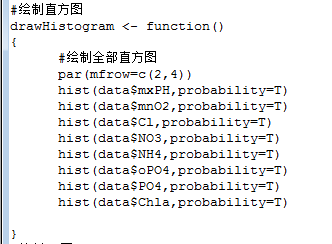
首先绘制mxPH的直方图



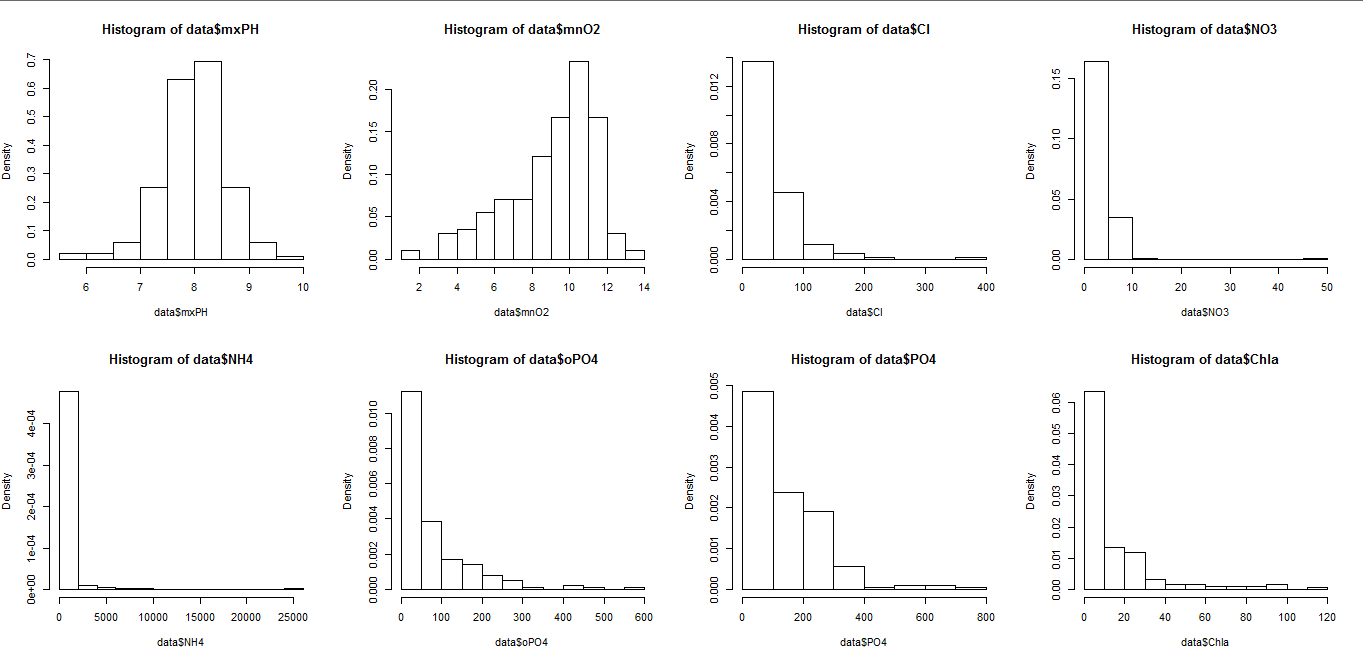
结果如下：



进而，我们绘制所有的直方图：



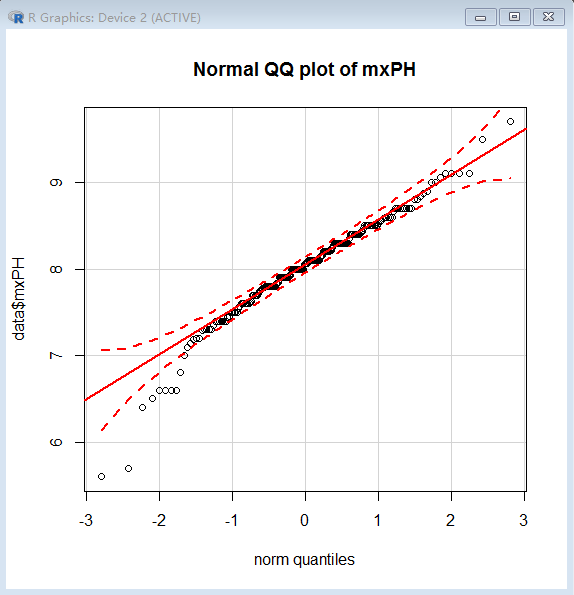
运行结果：



(2) 绘制QQ图

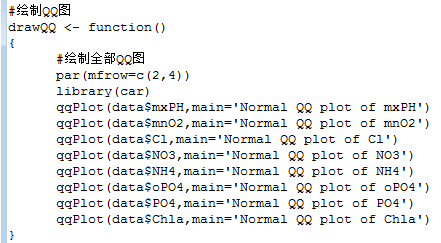
QQ图可以用来检验数据的分布是否为正态分布，我们使用R语言中的qqPlot()函数。

绘制mxPH的QQ图

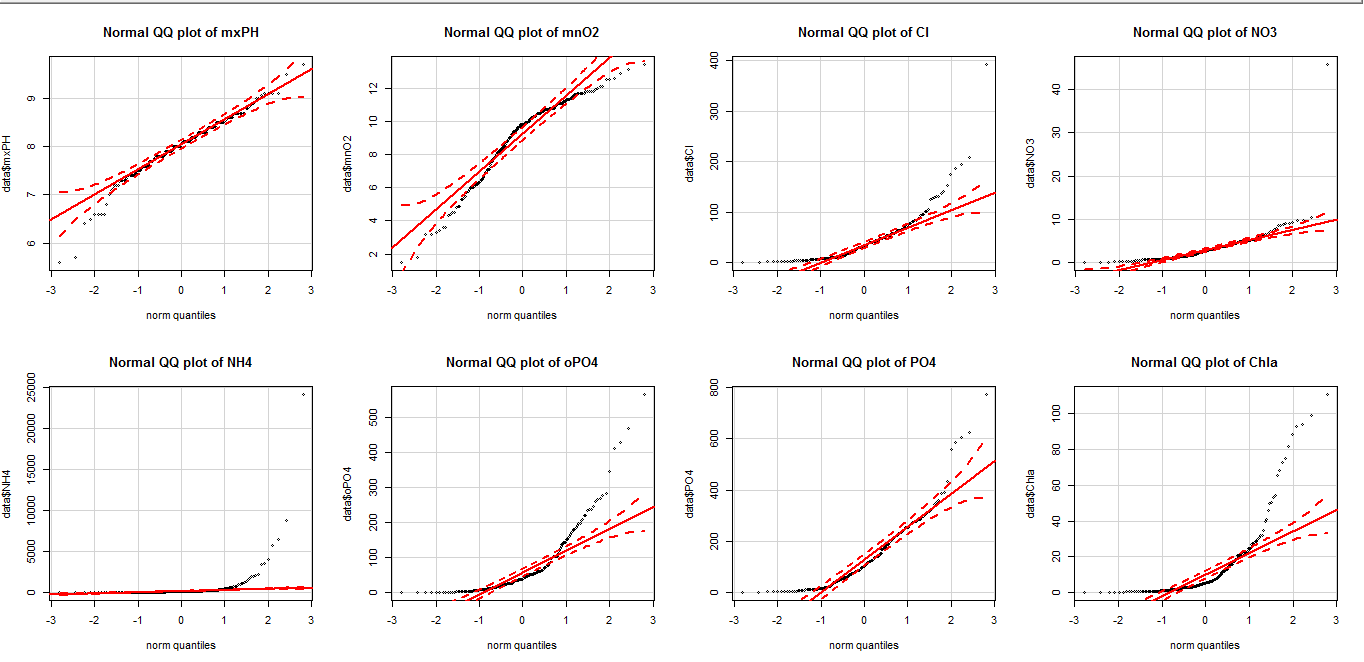


从QQ图可以看出变量有几个小的值明显在95%置信区间之外，因此不符合正态分布。

进而，我们绘制了全部的QQ图。

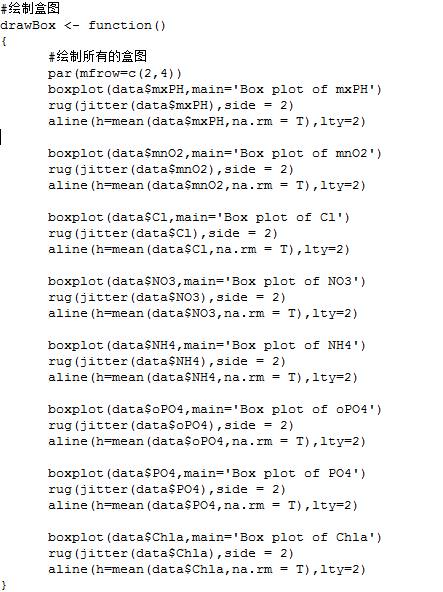


运行结果：

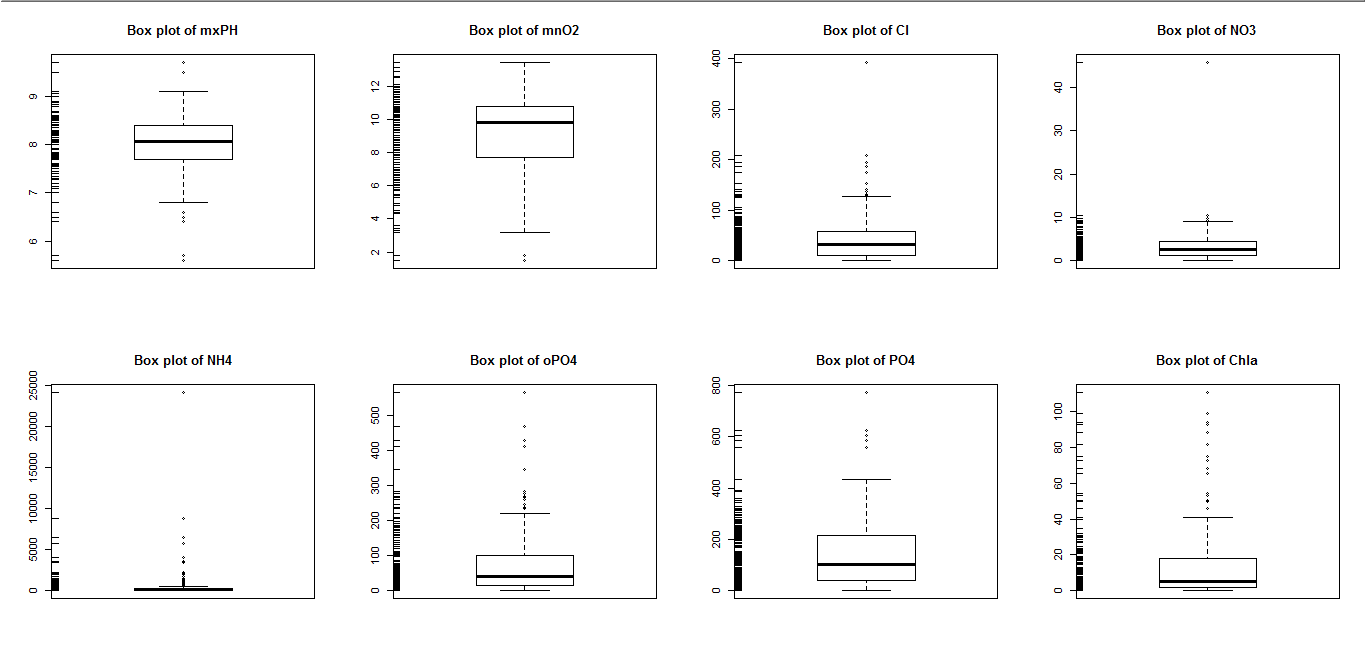


(3) 绘制盒图

调用boxplot函数。



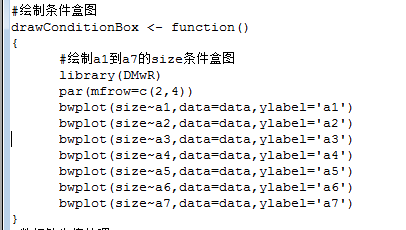
结果如图：



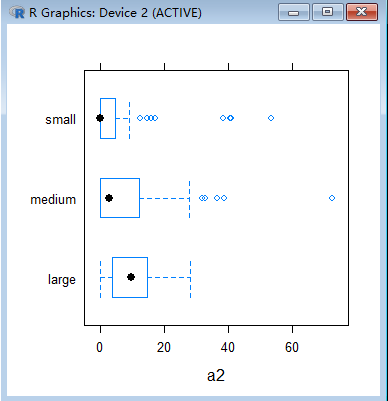
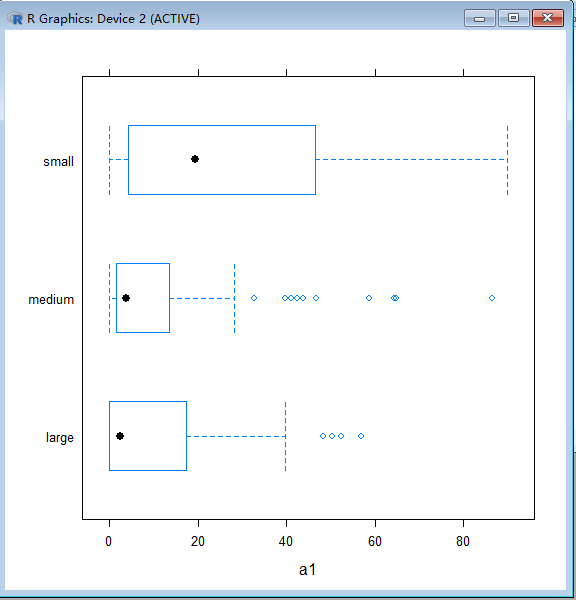
盒图上方小横线上面的小圆圈表示与其他值比较特别大的值，通常认为是离群值。与X轴平行的直线，是变量的均值所在的位置，将均值线与中位线进行比较，可知变量的多个离群值使得作为变量中心的均值产生了扭曲。

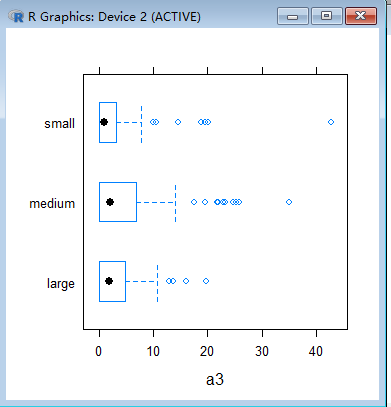
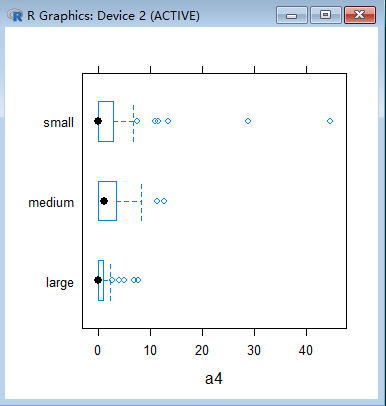
### 2.1.3 对7种海藻，绘制size条件盒图

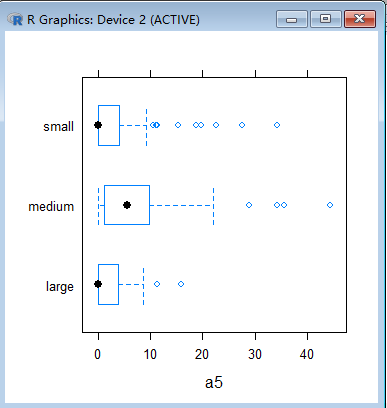
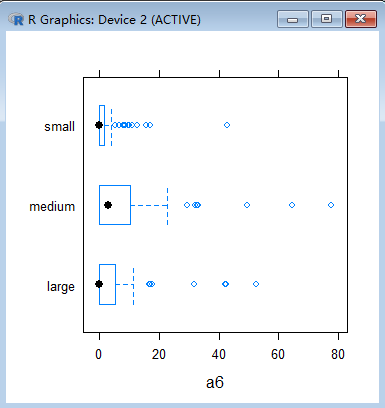
调用函数bwplot。

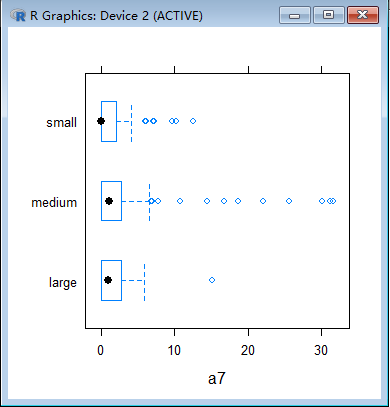


结果如图：





从上图中可以看出，在size为small的河中，a1频率最大；在size为medium河中，a5频率最大；在size为large的河中，a1的频率最大。

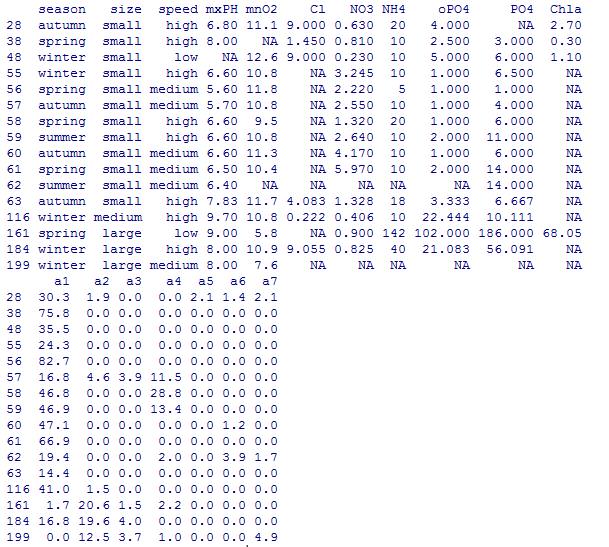
## 2.2 数据缺失的处理

### 2.2.1 剔除缺失部分

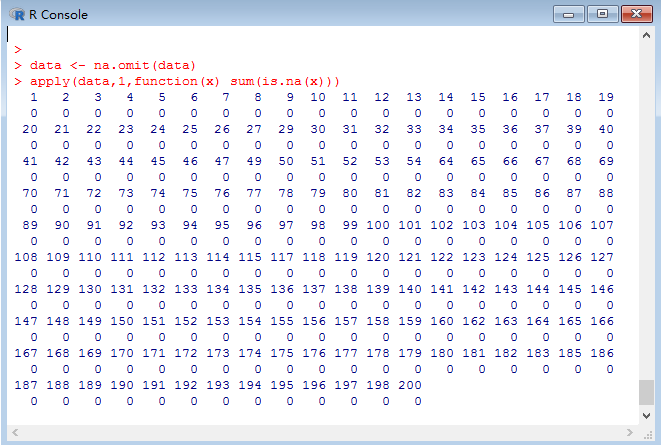
(1) 先查看缺失数据条目



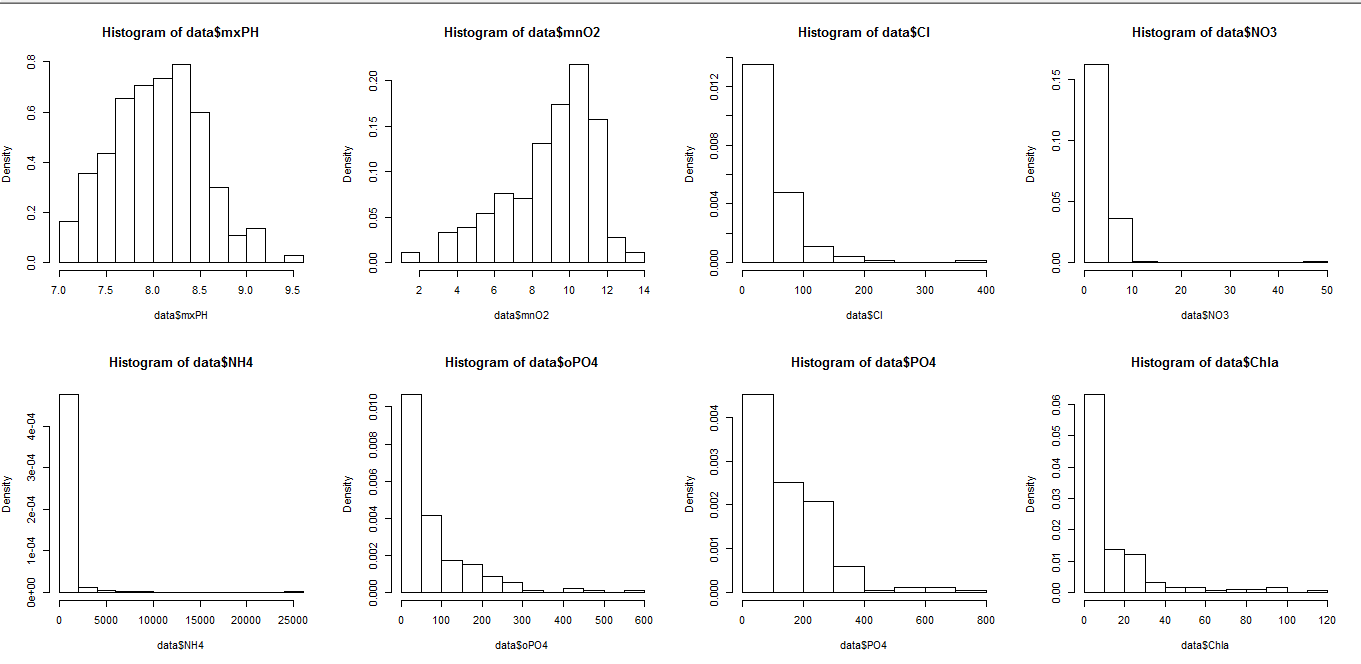
运行结果：

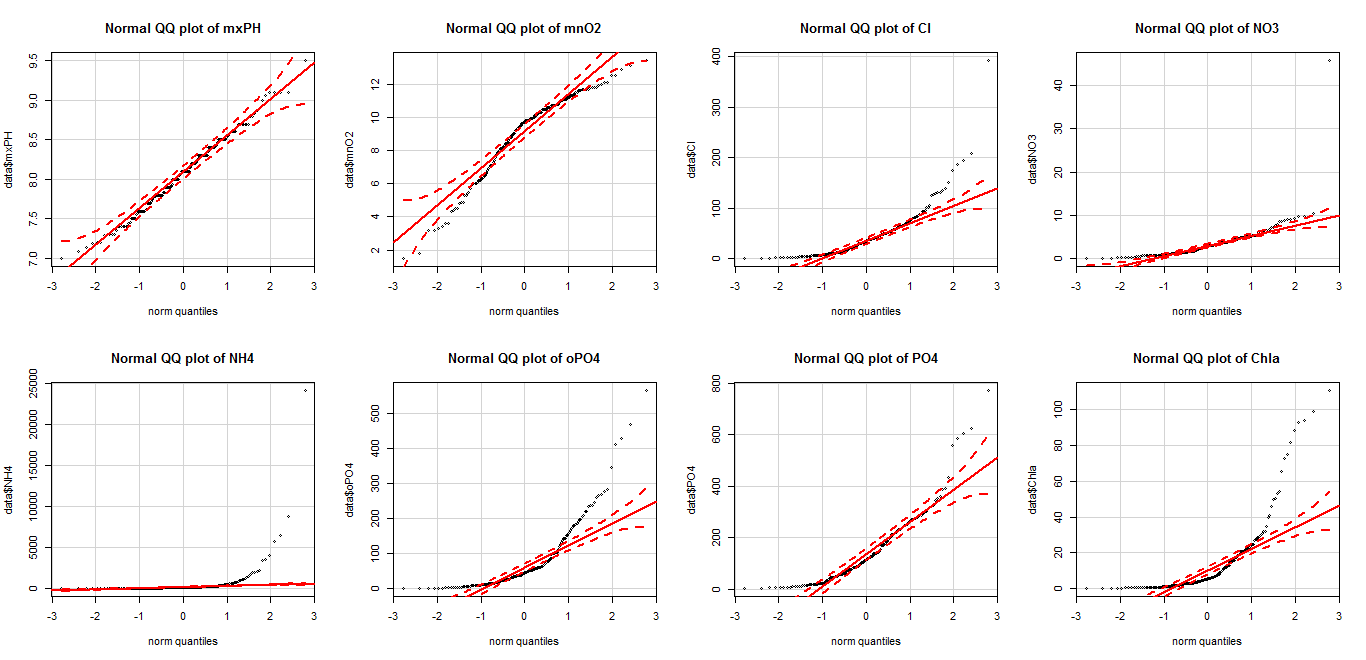


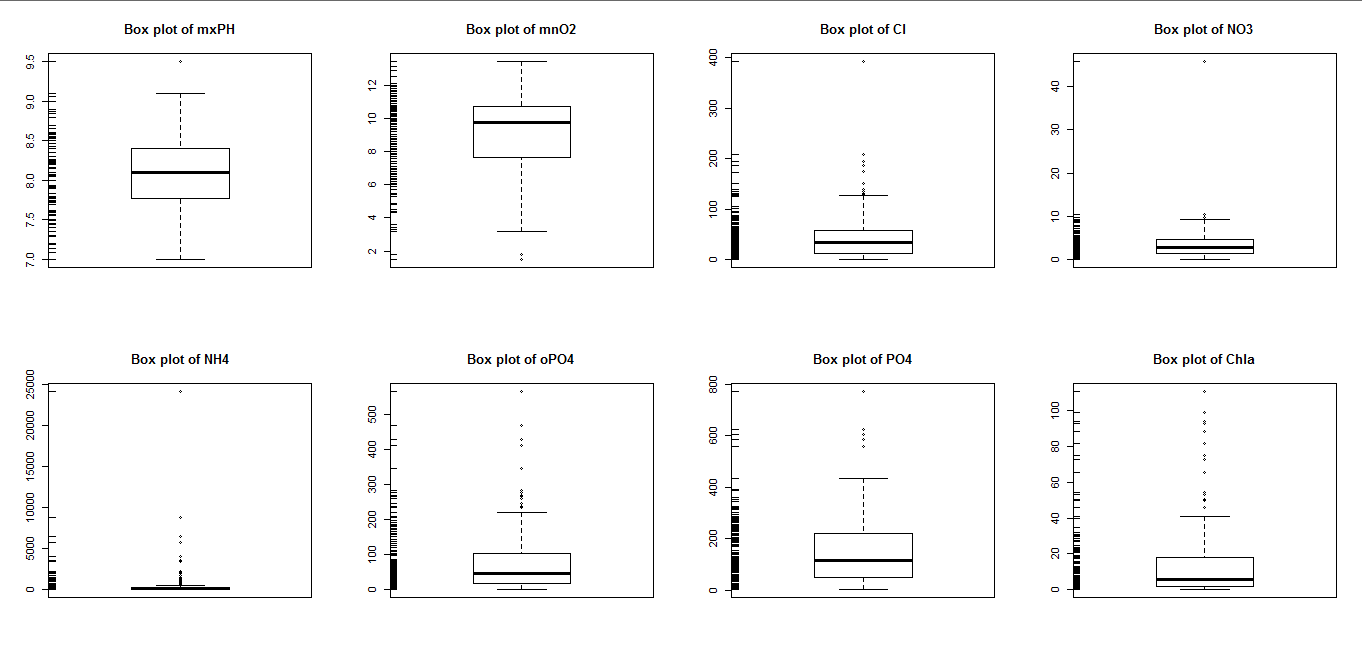
可以看出总共有16条数据，剔除掉之后影响也不大，因此剔除。



剔除数据之后，观察新得到数据集的直方图、QQ图、盒图。



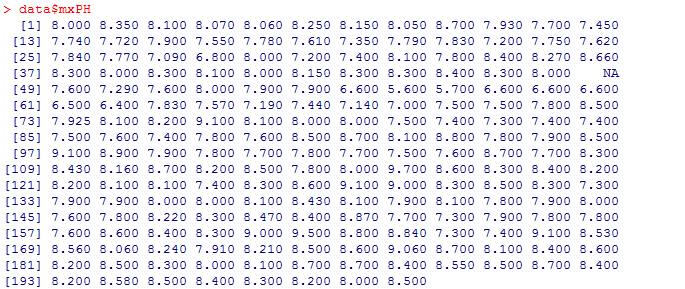




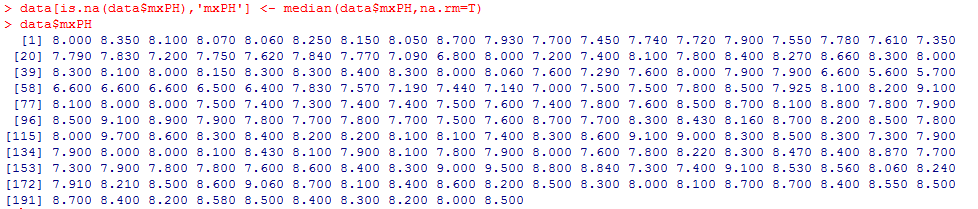
## 2.2.2 用最高频率值来填补缺失值

在这部分，我们用中位数来填补缺失值。

1. 首先显示原始mxPH数据

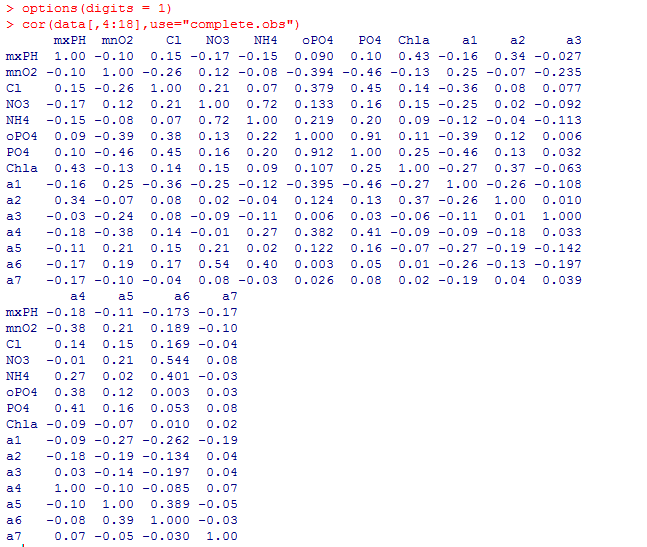


1. 将中位数填补缺失值

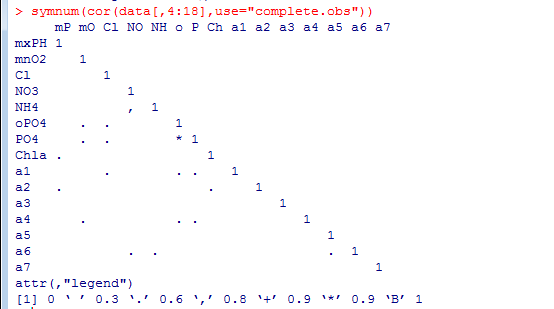


### 2.2.3 通过属性的相对关系来填补确实值

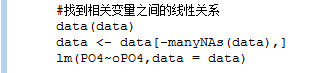
函数cor()可以产生属性之间的相关值矩阵，我们设为从4到18，因为前3个变量为标称变量，因此不予考虑。将参数设定为use=”complete.obs”，则R语言在计算相关值时不考虑缺失值。相关值为1时表示相应的两个变量之间有强正线性关系，相关值为-1时表示有负线性关系。之后R可以得到变量间线性相关的近似函数形式，可以让我们通过一个变量的值计算另一个变量的值。



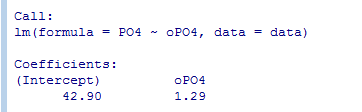
为了更清楚地表示，我们使用symnum函数来改善结果的输出形式。



从图中可以看出，变量PO4和变量oPO4之间相关度很高。因此可以相互填补这两个变量的缺失值。因此需要找到这两个变量之间的线性关系。



运行结果如下：



因此可得线性模型为PO4 = 1.29oPO4 + 42.90。即如果这两个变量不是同时有缺失值时，可以用这个公式计算变量的缺失值。我们将PO4与oPO4同时缺失的数据剔除后，还剩一个样本28在PO4上有缺失值。利用上述线性关系计算缺失值得填补值:



### 2.2.4 通过数据对象之间的相似性来填补缺失值

我们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最近的k个邻居。可以使用中位数来填补缺失值。

