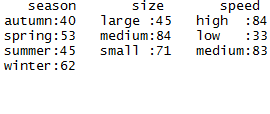
2120151020 罗伟

**海藻数据的分析报告**

数据分析基于R语言，IDE为RStudio环境。

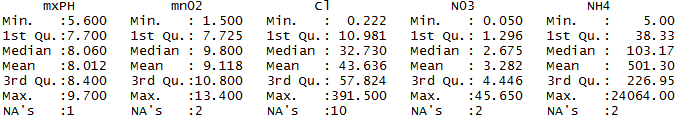
## 数据摘要

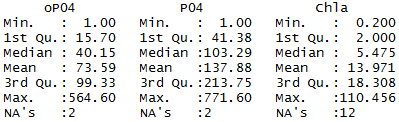
**1、标称属性频数**

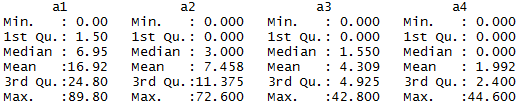


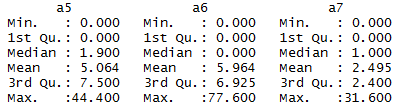
**2、数值属性各基本统计量情况：**

从上到下依次为最小值、一分位数、中位数、平均值、3分位数、最大值以及缺失值个数。









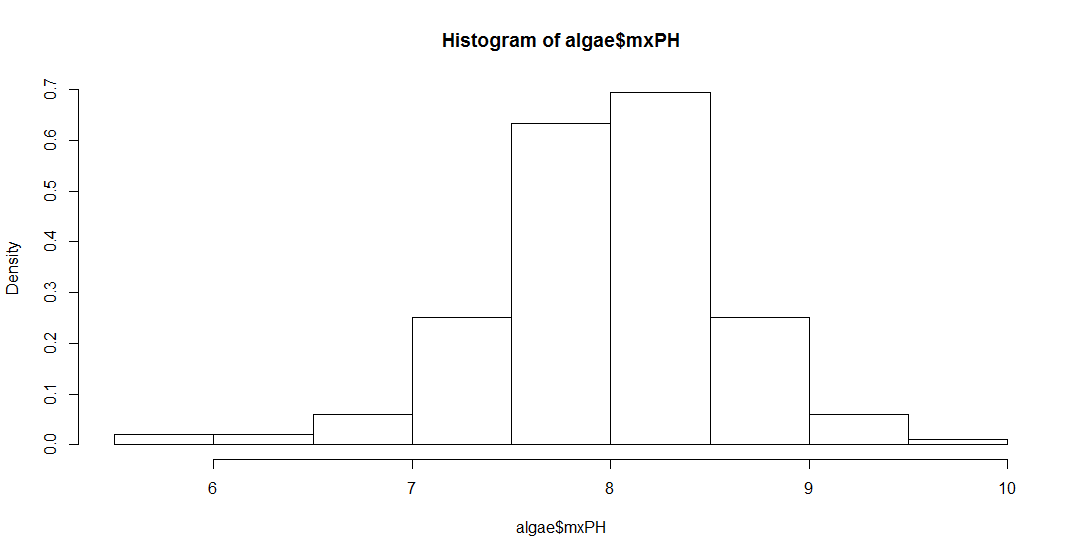
## 数据可视化

1. **绘制直方图并用q-q图检验分布是否为正态**

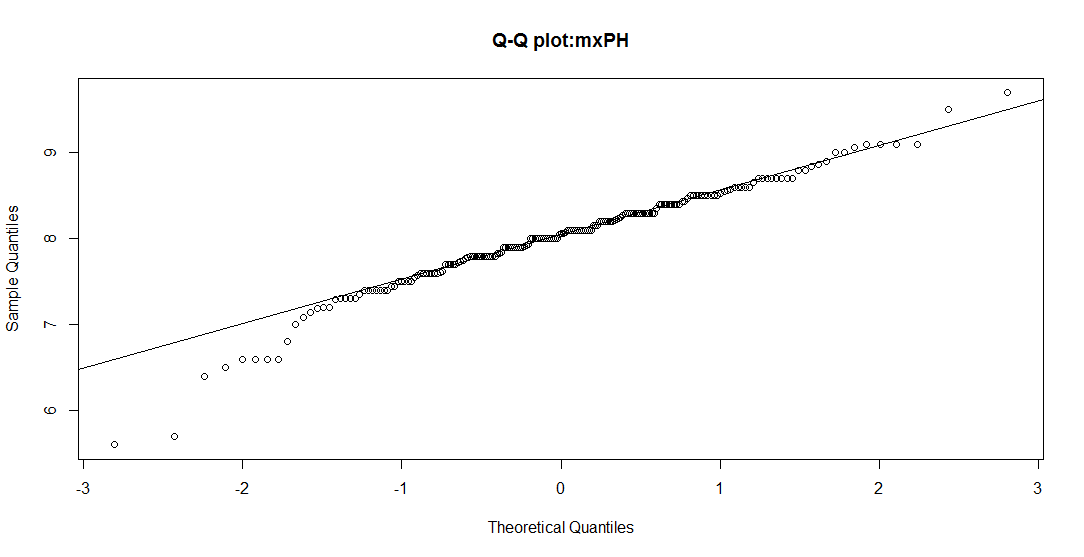
利用q-q图鉴别样本数据是否近似于正态分布,只需看QQ图上的点是否近似地在一条直线附近,而且该直线的斜率为标准差,截距为均值.

1. mxPH

直方图：



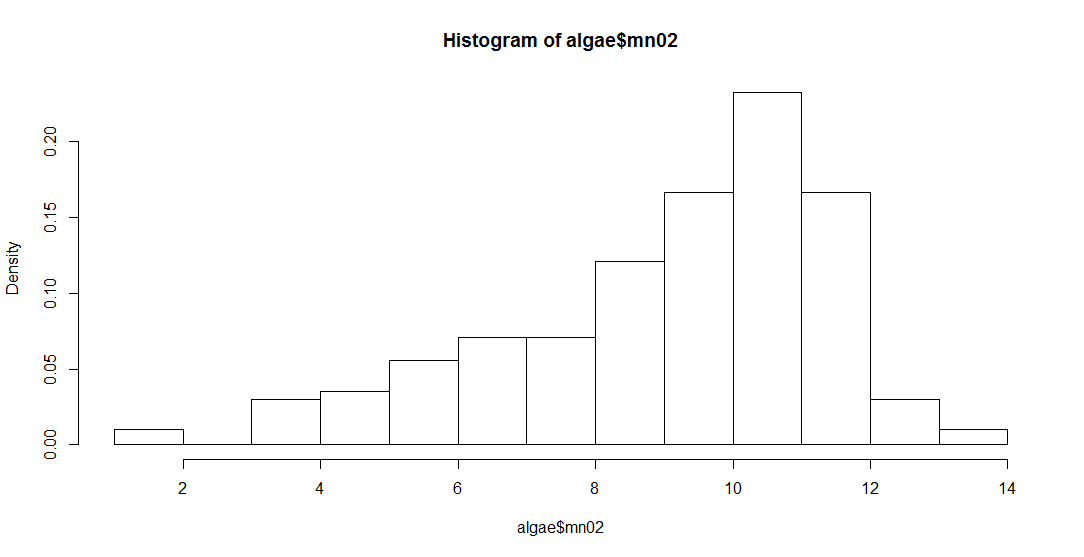
q-q图：



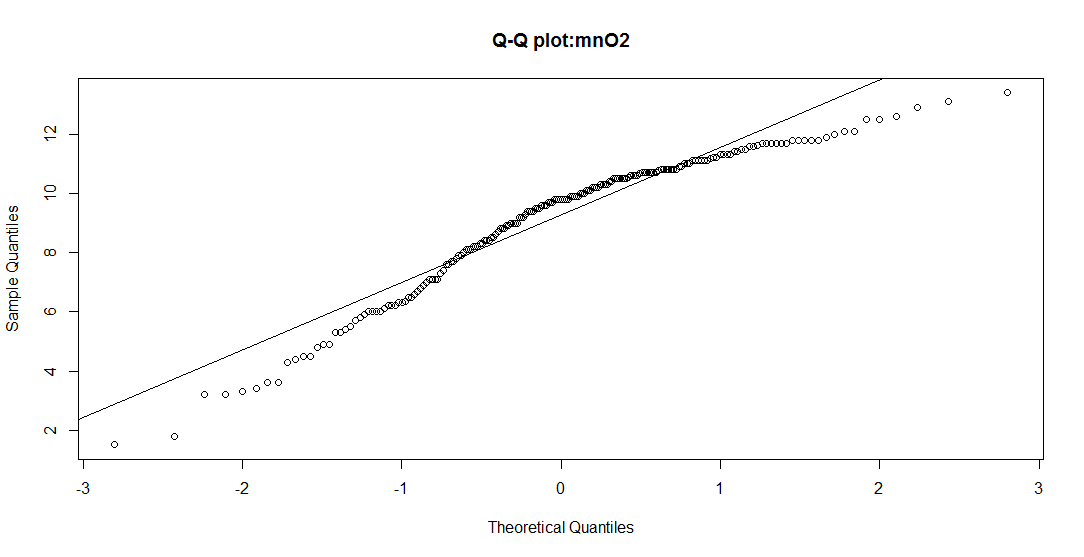
数据呈现为正态分布。

（2）mn02

直方图：



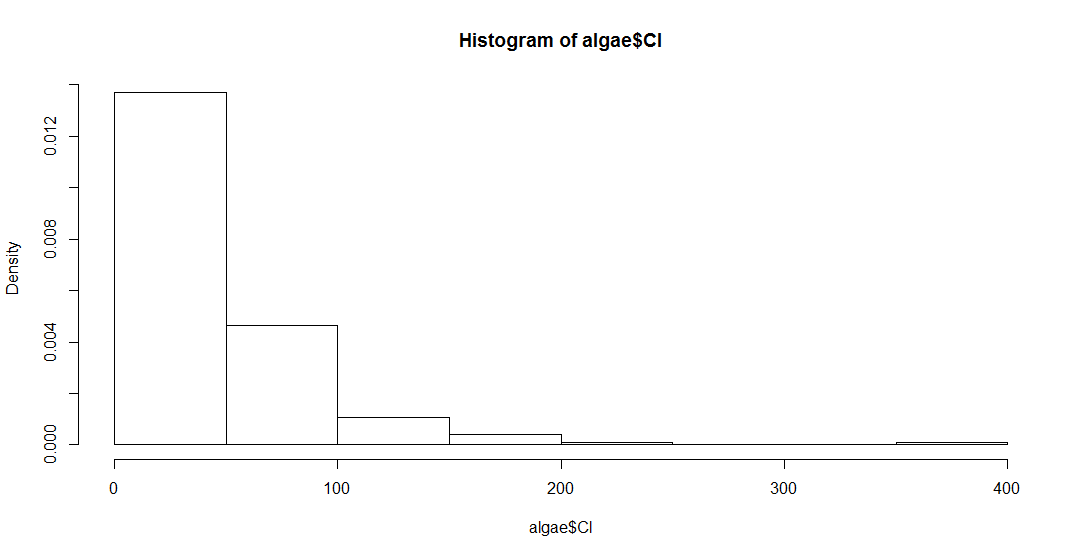
q-q图：



数据不为正态分布。

（3）CI

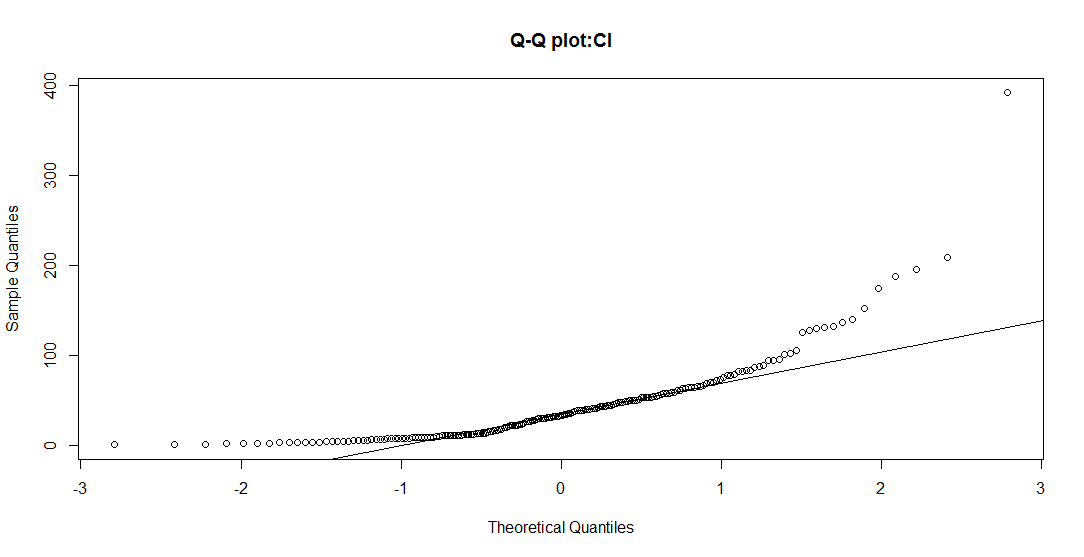
直方图：



这里可以看到，受离群值影响，直方图中对于正态分布的表现不明显，但实际上如果0-100的区间再进行划分还是能看出正态分布的效果的。进一步需要对数据进行处理后再进行作图分析，不过对于q-q图的绘图效果影响不大。

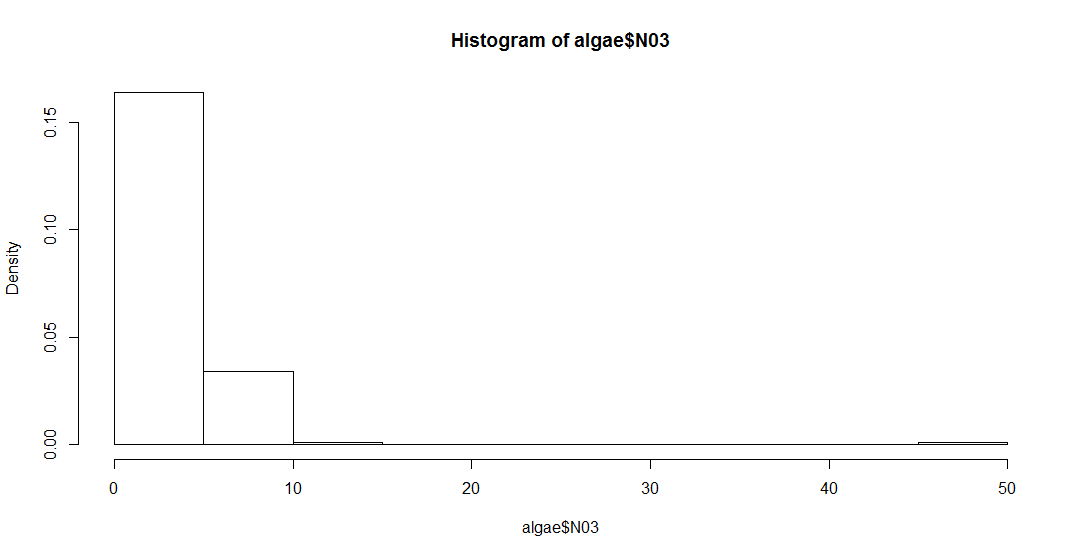
之后出现的几个数据直方图不呈现正态分布而q-q图表明是正态分布的原因也是因为离群值影响。

q-q图：

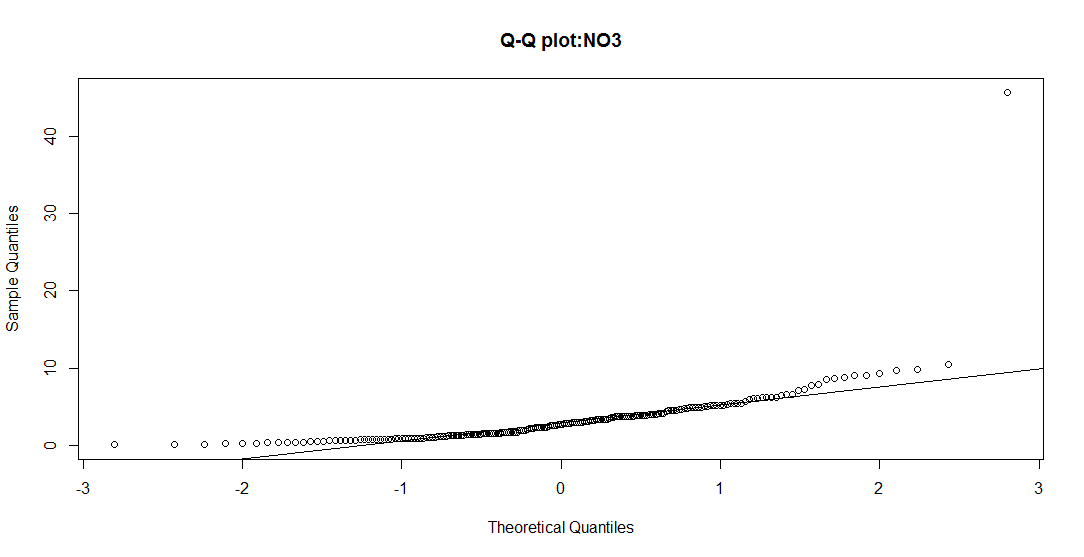


（4）NO3

直方图：



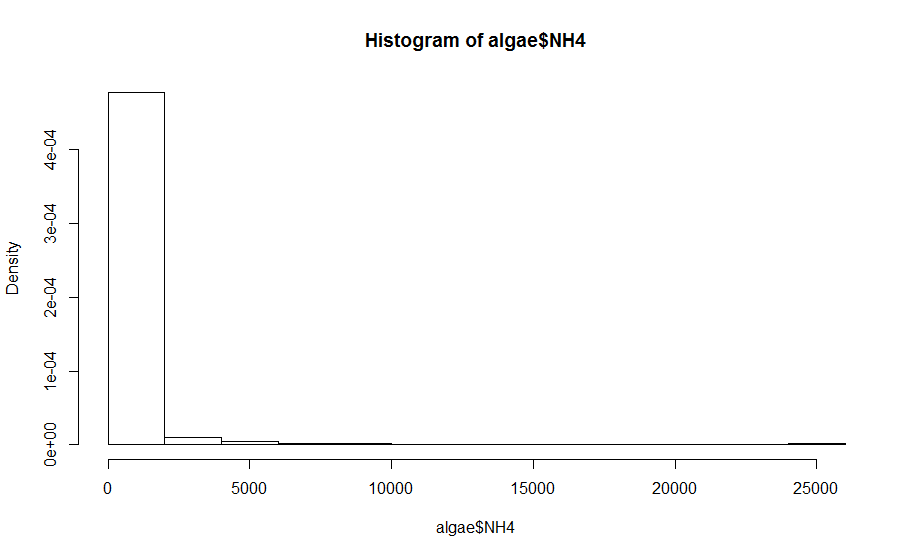
q-q图：



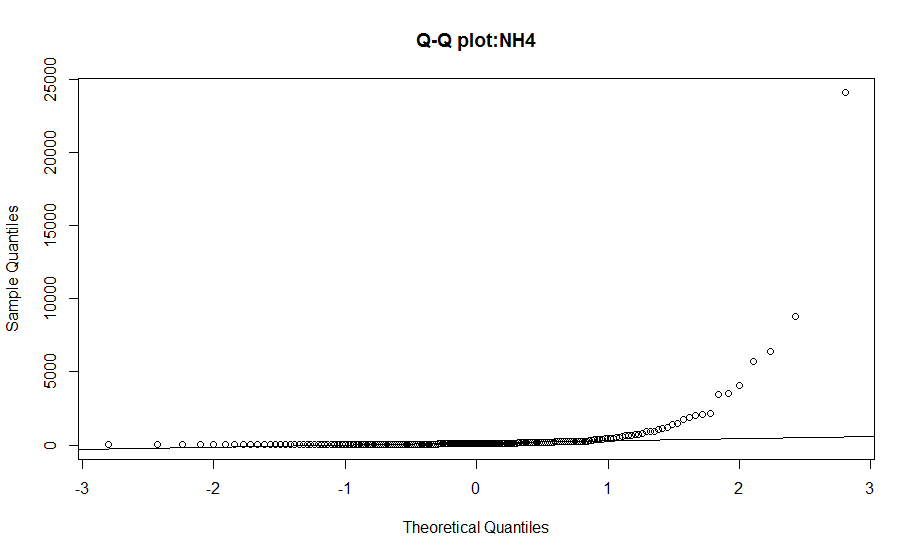
数据呈现为正态分布。

（5）NH4

直方图：



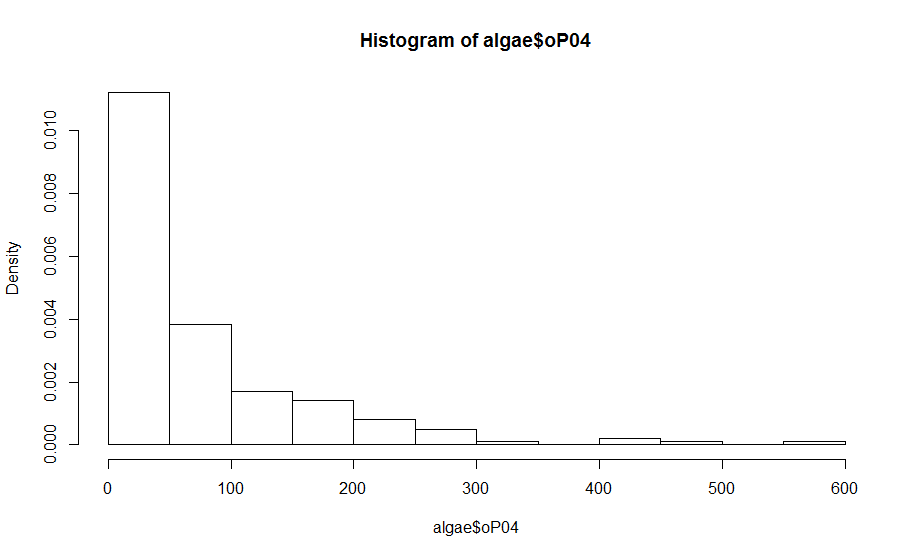
q-q图：



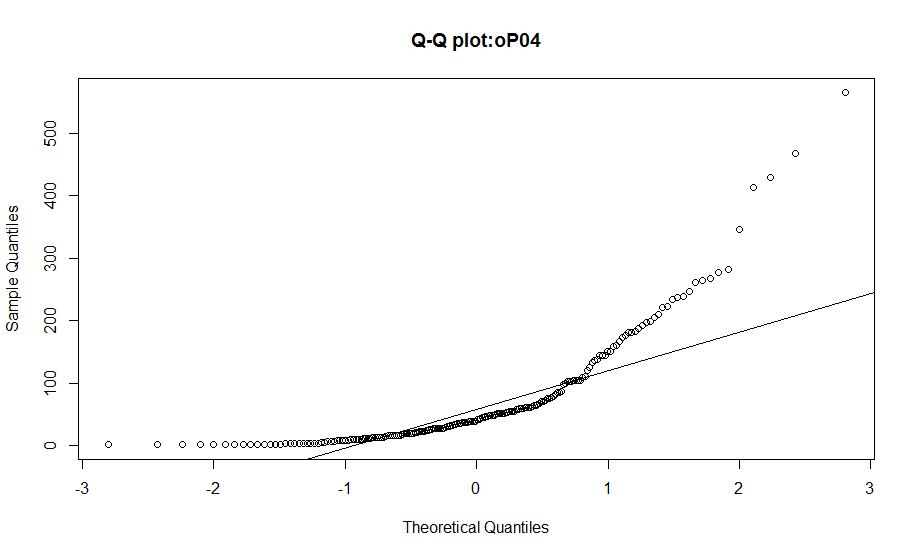
数据可以认为呈现正态分布。

（6）oP04

直方图：



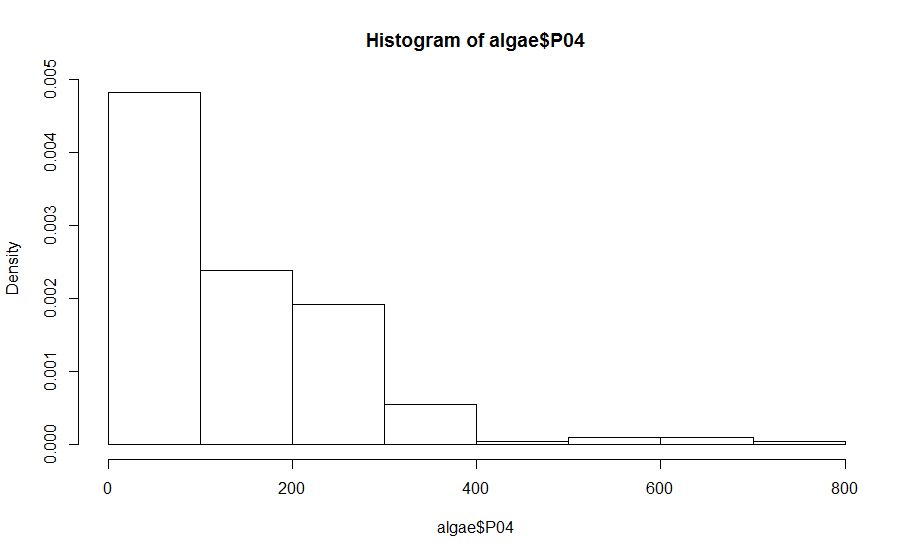
q-q图：



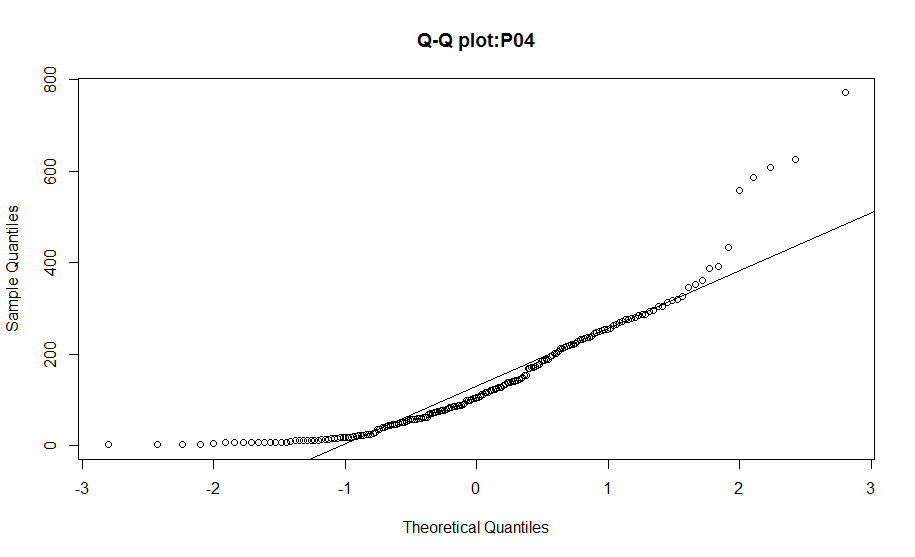
并不呈现正态分布。

（7）P04

直方图：



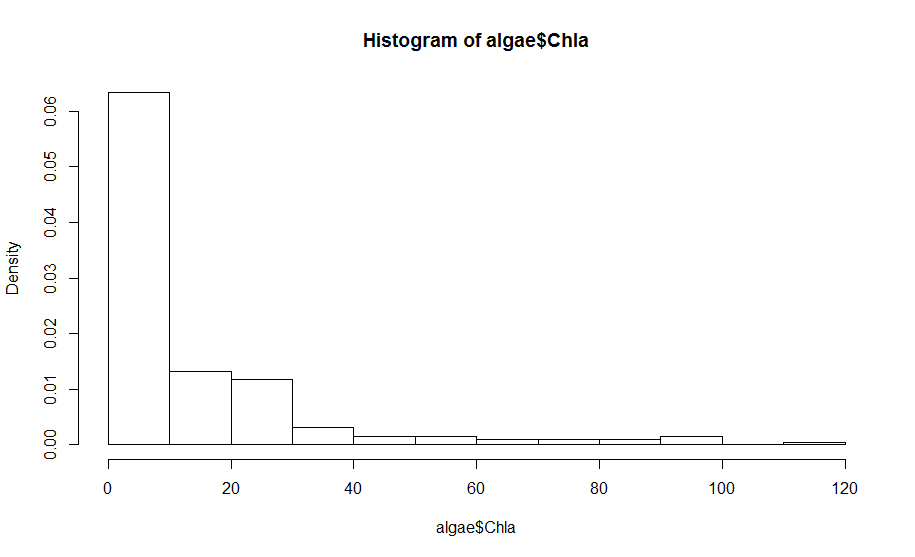
q-q图：



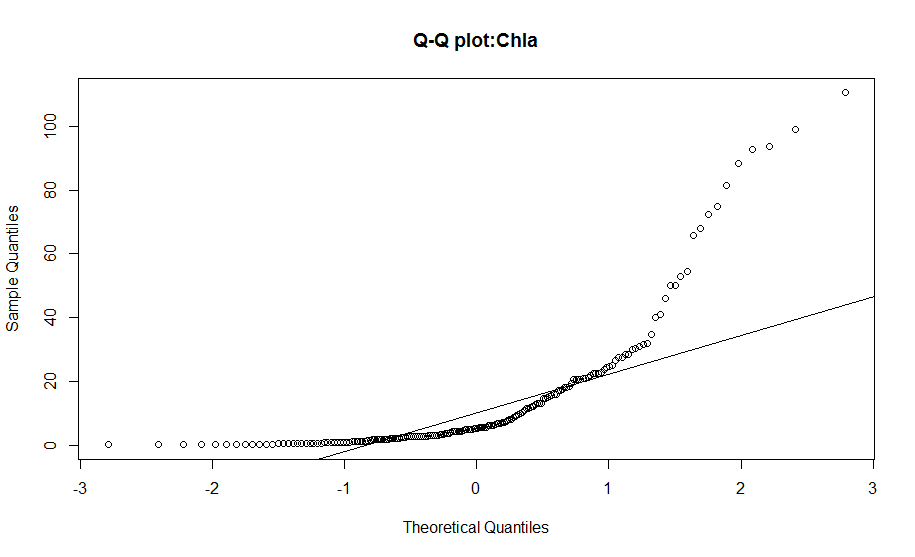
数据明显不为正态分布。

（8）Chla

直方图：



q-q图：

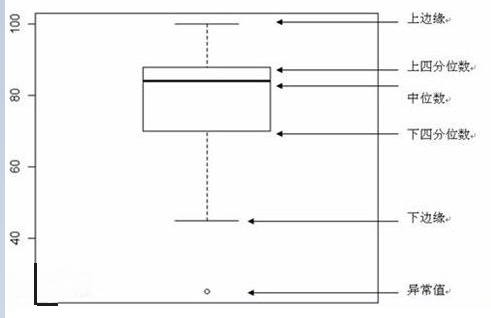


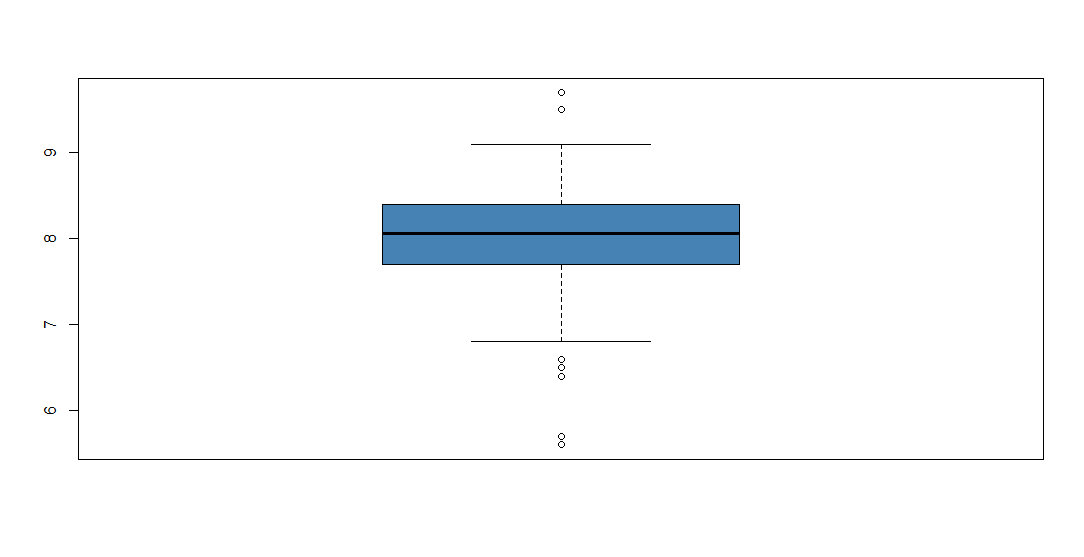
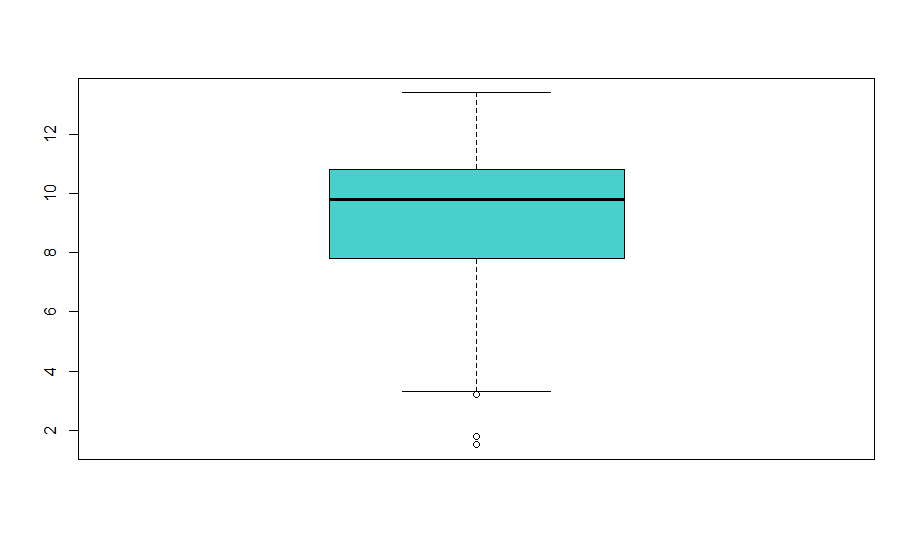
数据明显不呈现为正态分布。

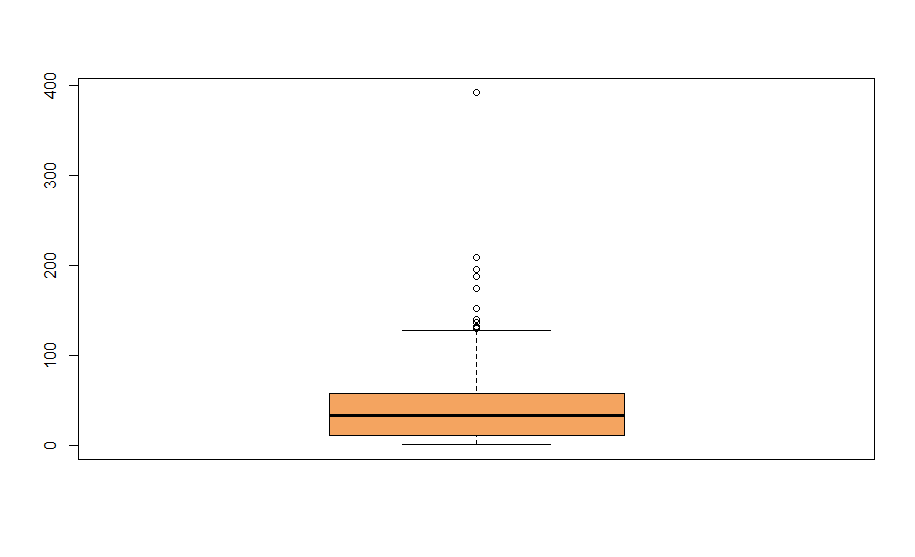
1. **绘制盒图，对离群值进行识别**

箱形图（Box-plot）又称为盒须图、盒式图或箱线图，是一种用作显示一组数据分散情况资料的[统计图](http://baike.baidu.com/view/222815.htm)。因型状如箱子而得名。

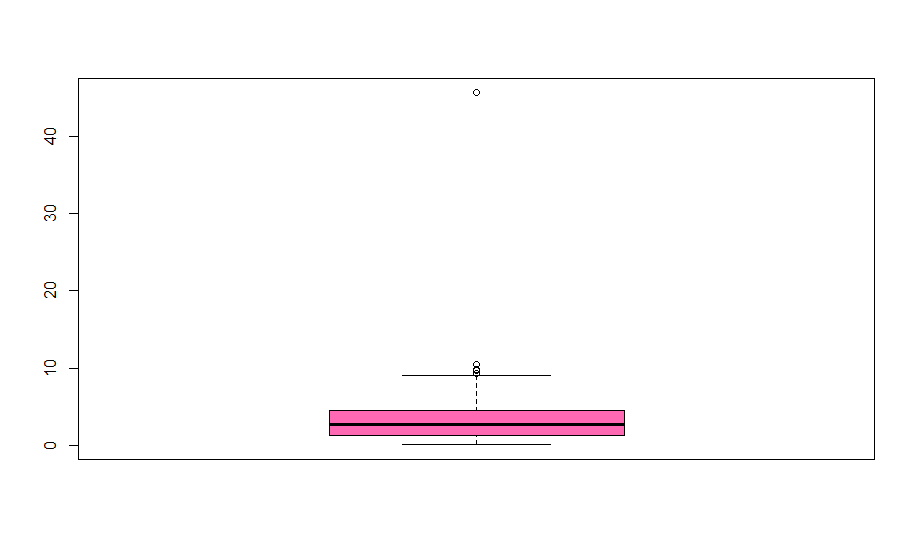
主要包含六个数据节点，将一组数据从大到小排列，分别计算出他的上边缘，上[四分位数](http://baike.baidu.com/view/999573.htm)，[中位数](http://baike.baidu.com/view/170892.htm)，下四分位数，下边缘，还有一个异常值。



1. mxPH 
2. mn02 

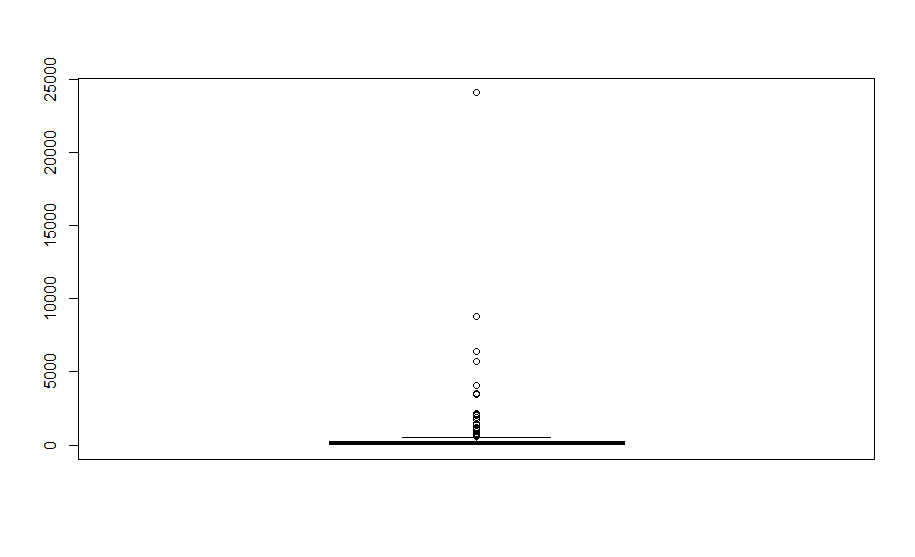
Cl 最上方点严重异常，为第153个数据第六列。



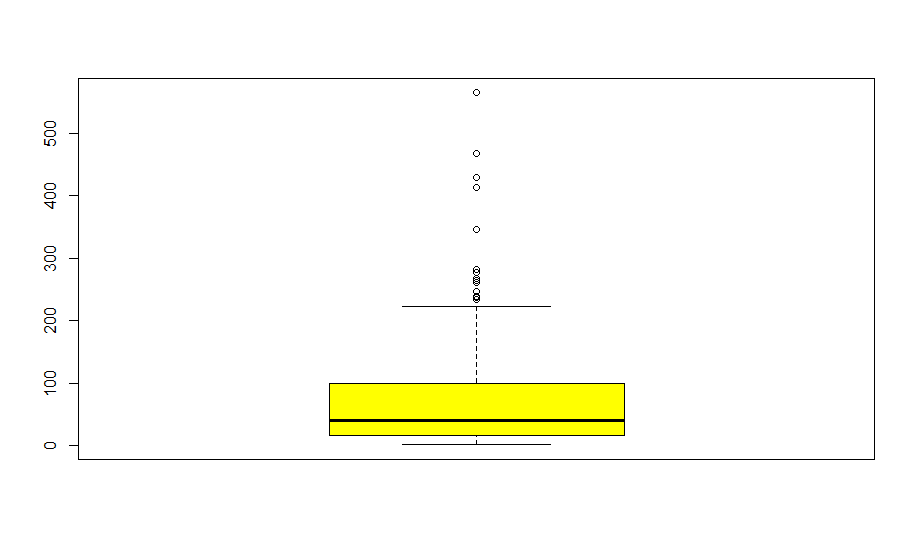
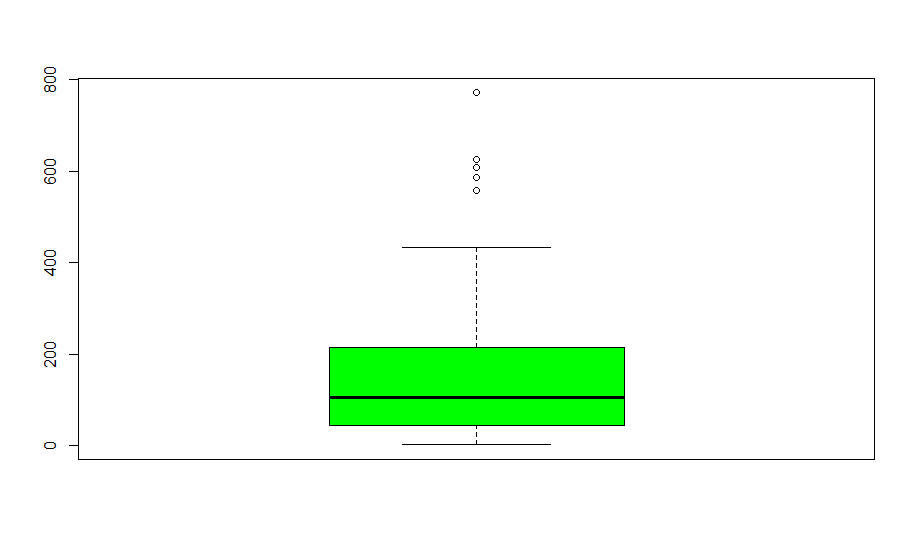
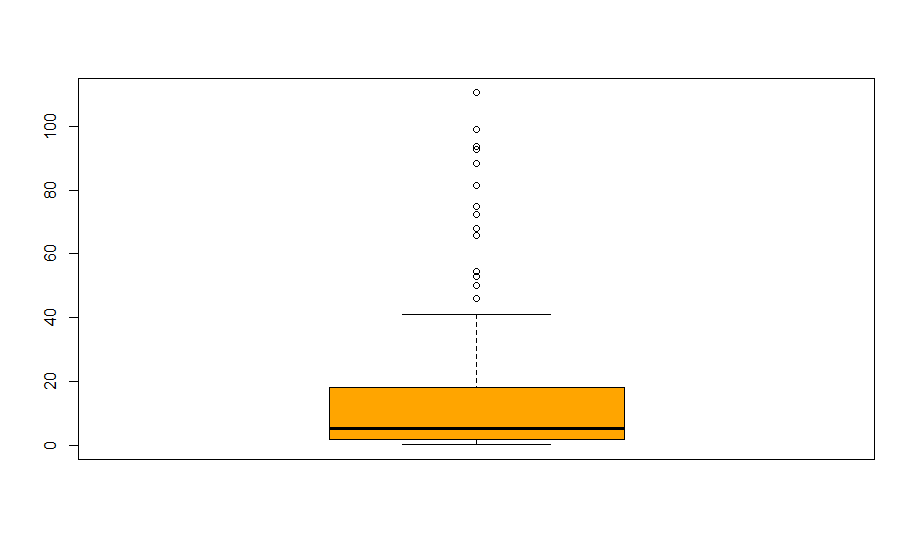
1. N03 

最上方点严重异常，为第153个数据第七列。



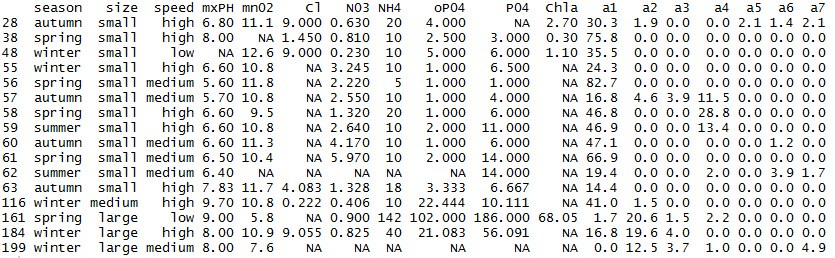
1. NH4 最上方点严重异常，为第153个数据第八列。



1. oP04 
2. P04 
3. Chla 

## 数据缺失的处理

首先找出有数据缺失的样本

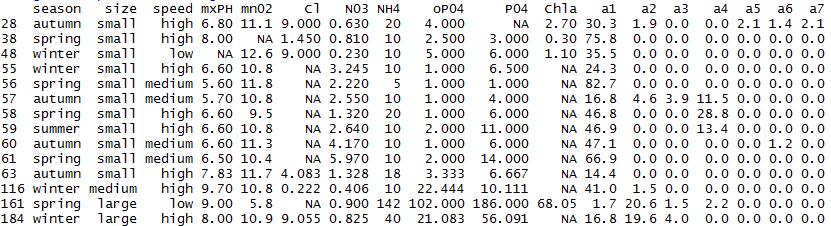


**对于不同的情况，应该对缺失进行不同的处理**

1. 将缺失部分剔除

对于62和199号样本来说，缺失的数据过多，本身已经失去了意义，故进行舍弃。

舍弃后的数据如下：



1. 用最高频率值来填补缺失值

统计出所有存在缺失属性中频数最大的值。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 项目 | mxPH | mnO2 | CI | NO3 | NH4 | oPO4 | PO4 | Chla |
| 值 | 8.3 | 9.8 | 7 | 0.23 | 10 | 1 | 6 | 0.8 |
| 频数 | 14 | 8 | 3 | 2 | 9 | 5 | 4 | 8 |

可以看到，最高频数对应值对于浮点数数据来说并不能很好地代表总体的情况，大多频数都很小，而且对应值不具有代表性，大多与中位数和均值都相距甚远，所以最终的数据预处理中并没有采用此方法。

1. 通过属性的相关关系来填补缺失值

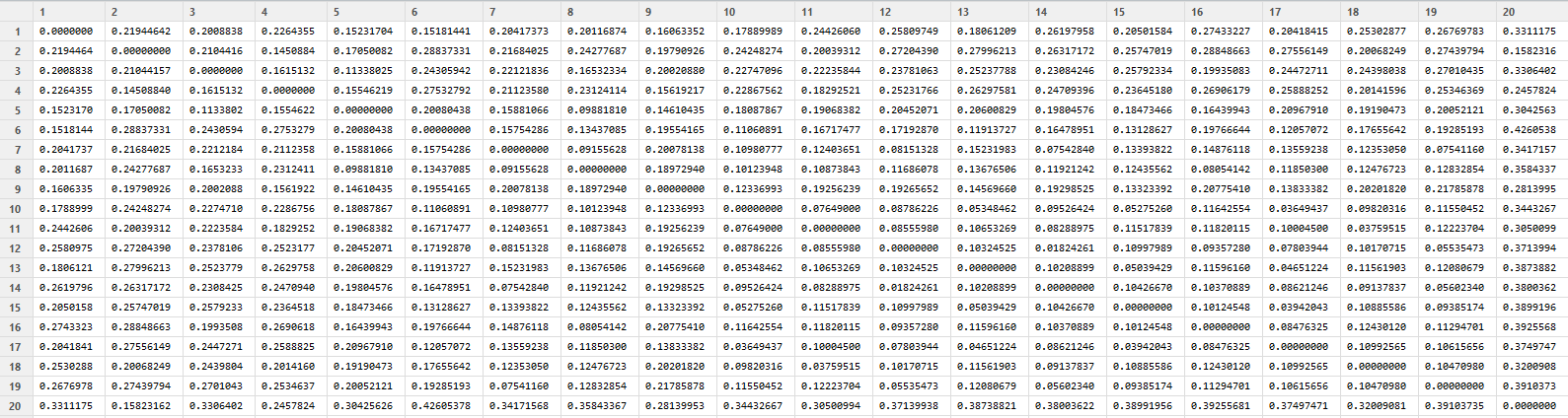
由于缺乏背景知识，对于属性间的相关关系不明了，故此部分没能完成。

1. 通过数据对象之间的相似型来填补缺失值

**要找到相似的样本，就需要样本定义之间的距离。**

**R语言提供的cluster包里有daisy函数，能计算具有混合变量样本间的距离。**

**将计算出的所有样本间的距离保存为一个200x200的矩阵**



**之后对存在缺失的样本数据r的缺失属性na进行填补处理：**

1. **取出与r最相似的10个样本**
2. **对10个样本对应的na属性计算中位数**

**最终得到填补的数据主要是进行了缺失部分剔除以及通过相似型属性填补的步骤。对于异常点由于定义尚不明确，故只是在文档中指出了一些特别异常的数据，而并没有对原数据进行替换或者删除处理。**

**最终得到的数据共198条。**