# 海藻数据的分析

韩学博 2120151008

## 1. 问题描述

某些高浓度的有害藻类对河流生态环境的破坏是一个严重的问题。它们不仅破坏河流的生物，也破坏水质。能够监测并在早期对海藻的繁殖进行预测对提高河流质量是很有必要的。

针对这一问题的预测目标，在大约一年的时间内，在不同时间内收集了欧洲多条河流的水样。对于每个水样，测定了它们的不同化学性质以及7种有害藻类的存在频率。在水样收集过程中，也记录了一些其他特性，如收集的季节、河流大小和水流速度。

## 2. 数据说明

**下载数据:** [地址](http://pan.baidu.com/s/1hsIwMBa)

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。

## 3、数据分析要求

### 3.1 数据可视化和摘要

数据摘要

* 对标称属性，给出每个可能取值的频数，
* 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。

数据的可视化

针对数值属性，

* 绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。
* 绘制盒图，对离群值进行识别

对7种海藻，分别绘制其数量与标称变量，如size的条件盒图

### 3.2 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:

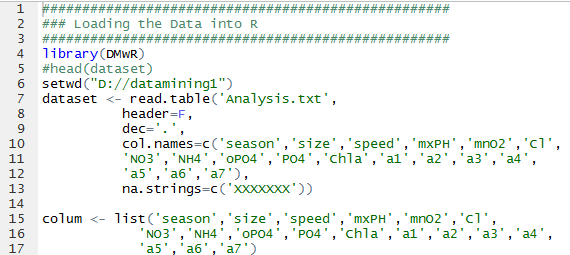
* 将缺失部分剔除
* 用最高频率值来填补缺失值
* 通过属性的相关关系来填补缺失值
* 通过数据对象之间的相似性来填补缺失值

处理后，可视化地对比新旧数据集。

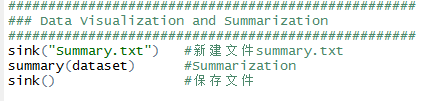
## 4、代码分析

### 读入数据

使用R语言自带的函数读入Analysis.txt里的数据。



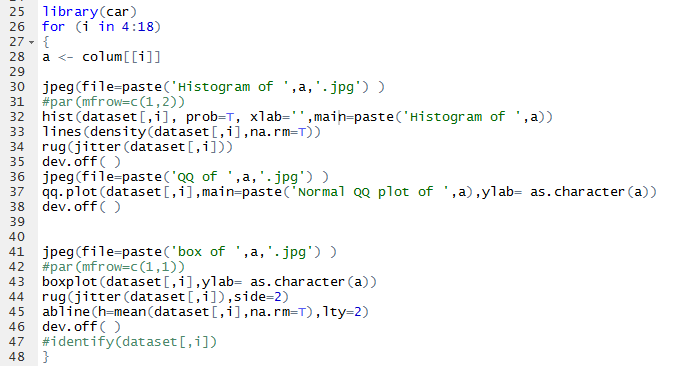
### 数据摘要



### 数据可视化

生成直方图、QQ图、盒图。

以生成直方图的代码为例：



## 数据分析

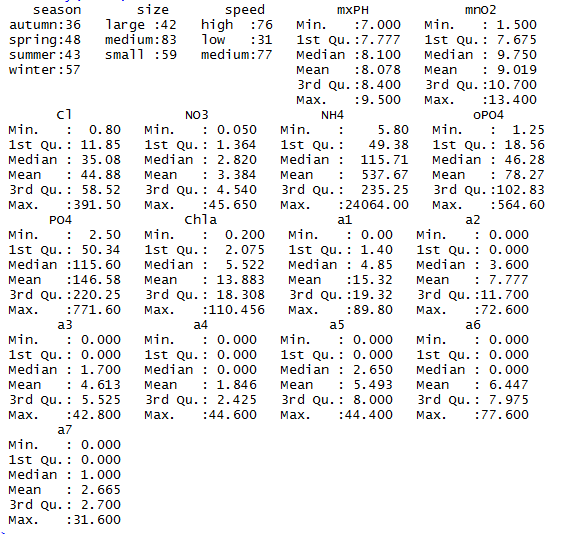
分别使用下列四种策略对缺失值进行处理:

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

处理后，可视化地对比新旧数据集。

### 将缺失数据剔除并可视化

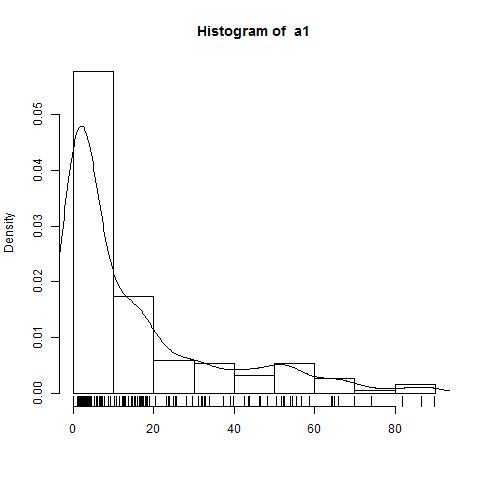
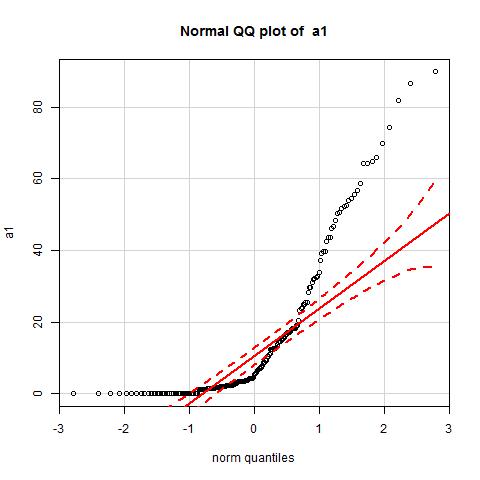
将缺失的数据直接剔除，处理后的摘要信息如下：



#### 使用直方图表示，使用QQ 图检验是否满足正态分布

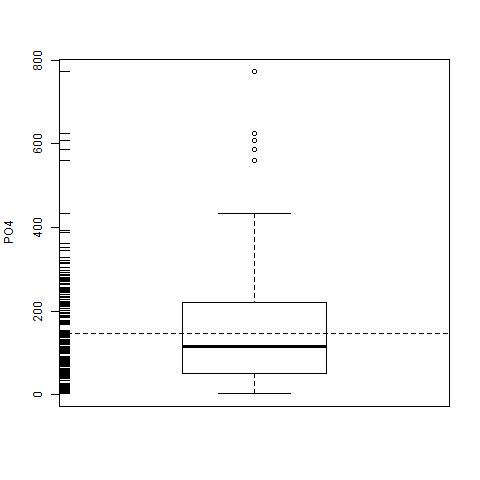
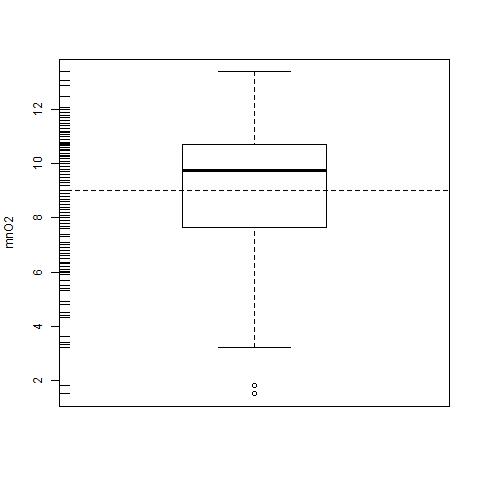
对于数值属性列，画出了所有的直方图和QQ图，下面选取a1 属性说明。

对于a1 属性，通过直方图和密度曲线可以看出不符合正态分布；使用QQ图验证可以验证这个结论。

#### 2) 画出数值属性的盒图

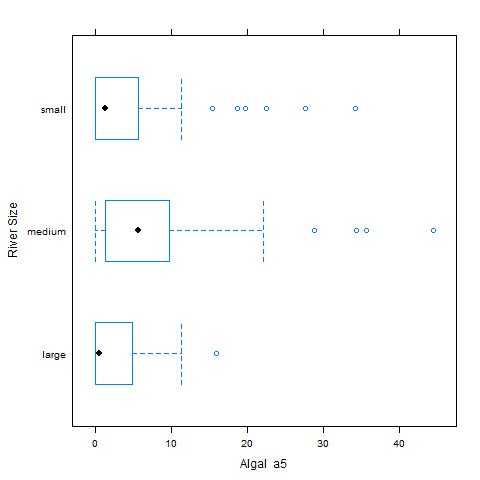
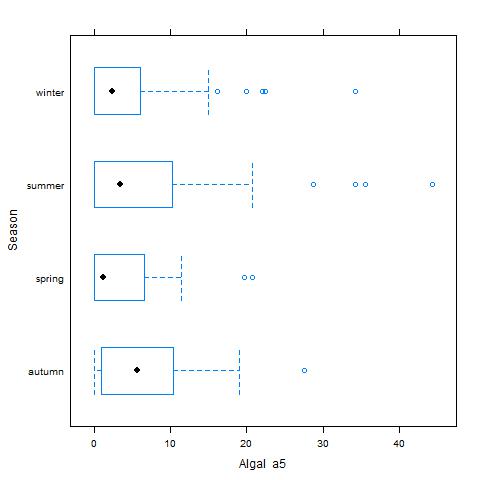
以mno2 和NH4 对应的数据为例，mn02有两个离群点，而NH4的离群点比较多。

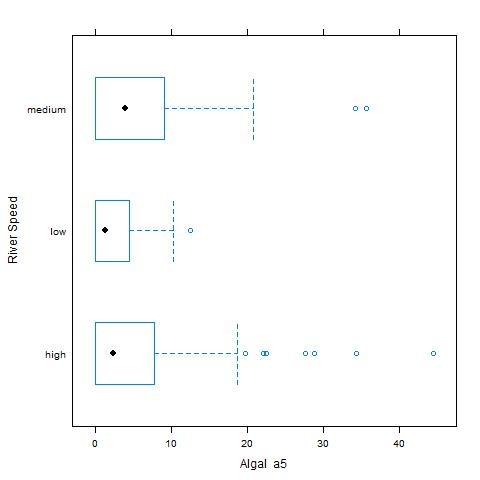


#### 3) 分别画出7种海藻和季节，河流大小，流速之间的条件盒图

下面选取a5属性为例。

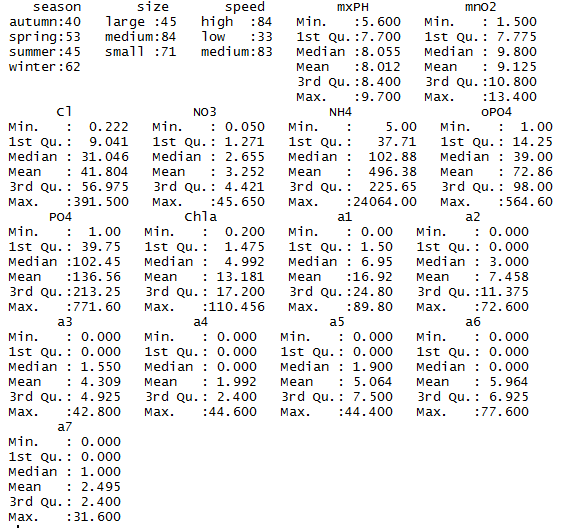
a 5 在夏天数量比较多，春天比较少，在中等河流中比较多，在中等流速的河流中数量多一些。





### (2) 用最高频率值来填补缺失值并可视化

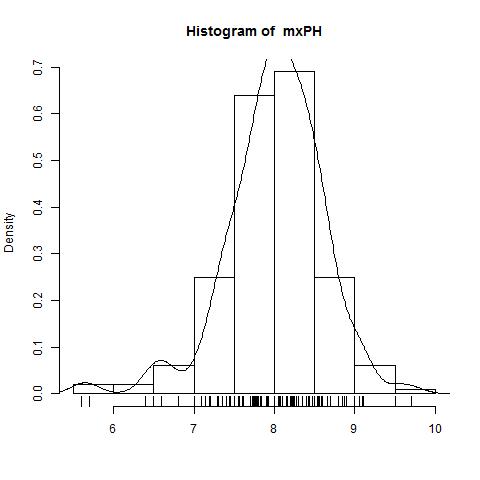
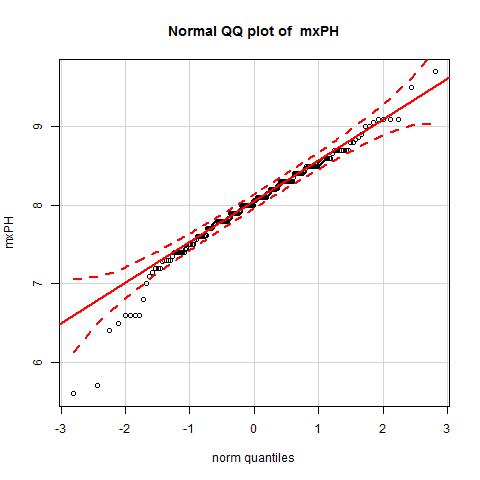
对于缺失样本采用使用最高频率的值来填补，处理结束后数据的摘要信息如下：



#### 使用直方图表示，使用QQ 图检验是否满足正态分布

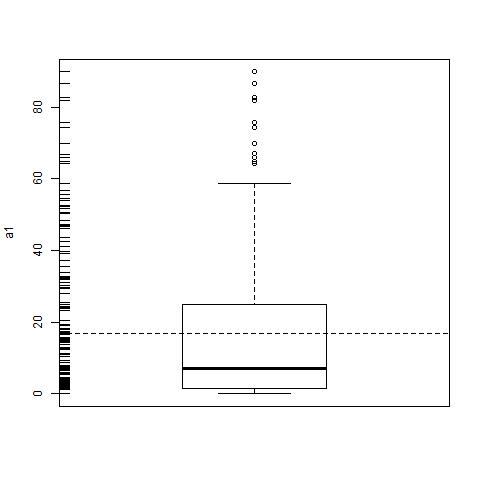
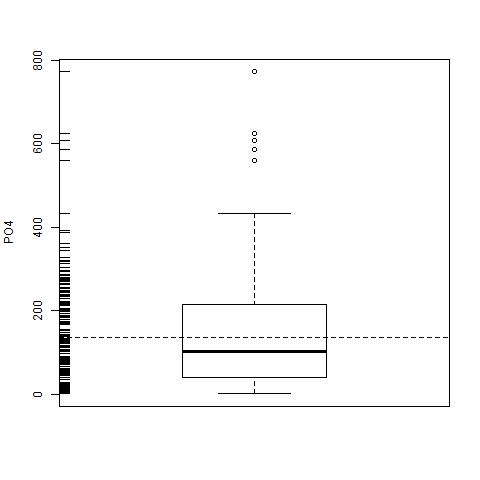
对于数值属性列，画出了所有的直方图和QQ图，下面选取mxPH属性说明。

对于mxPH属性，通过观察直方图和密度曲线发现符合正态分布；使用QQ图可以验证这个结论。

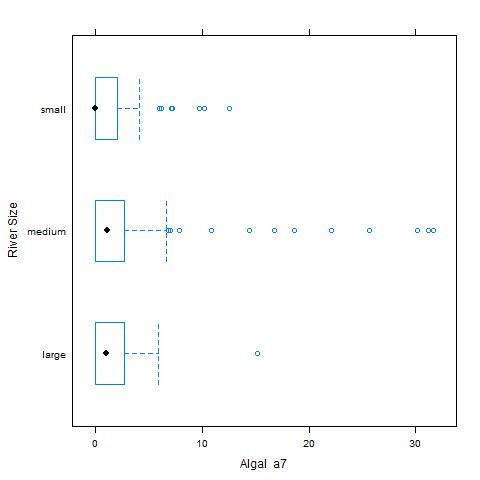
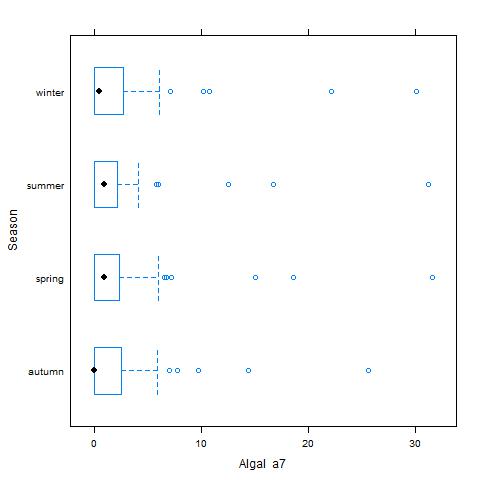
#### 2) 画出数值属性的盒图

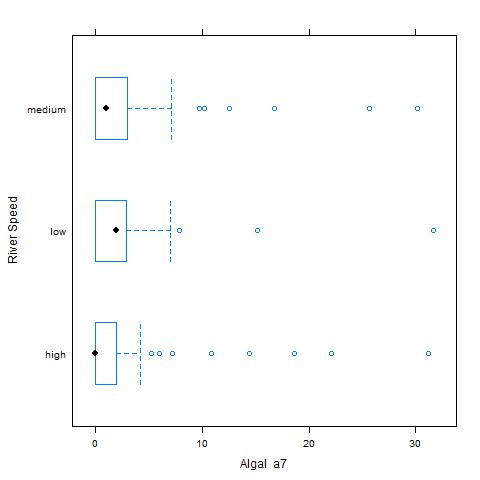
下面以PO4 和a1对应的数据为例，其他属性的在文件夹中，他们都存在离群点，可以看出都是偏大的离群点。



#### 3) 分别画出7种海藻和季节，河流大小，流速之间的条件盒图

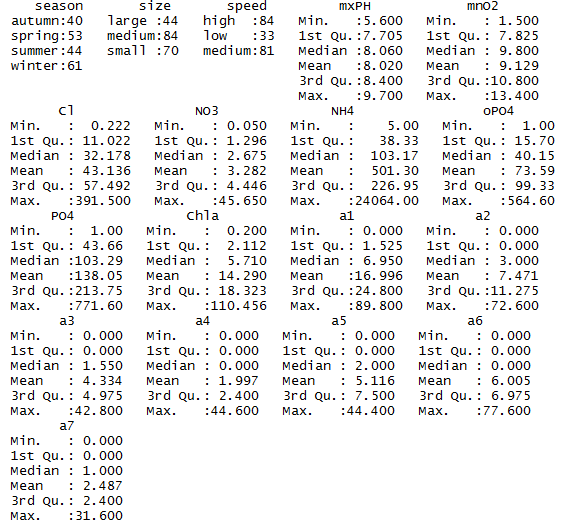
a 7 在春天数量比较多，在中等河流中比较多，在低流速的河流中数量多一些。





### (3) 通过属性的相关关系来填补缺失值并可视化

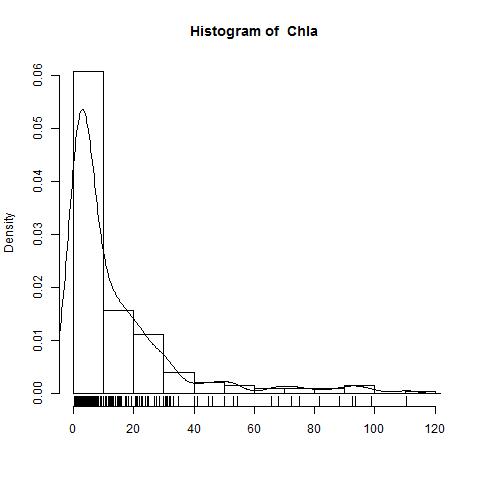
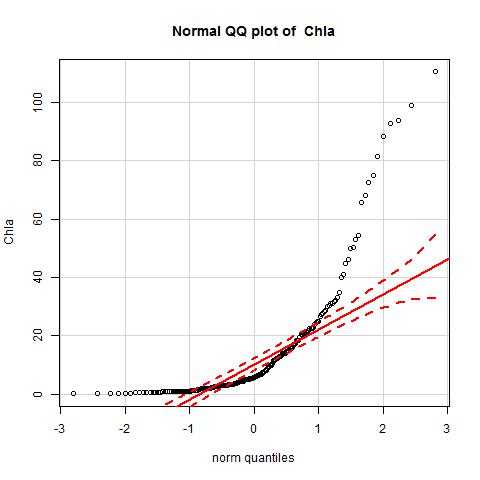
对于缺失样本我们采用采用属性的相关关系来填补，处理结束后数据的摘要信息如下：



#### 使用直方图表示，使用QQ 图检验是否满足正态分布

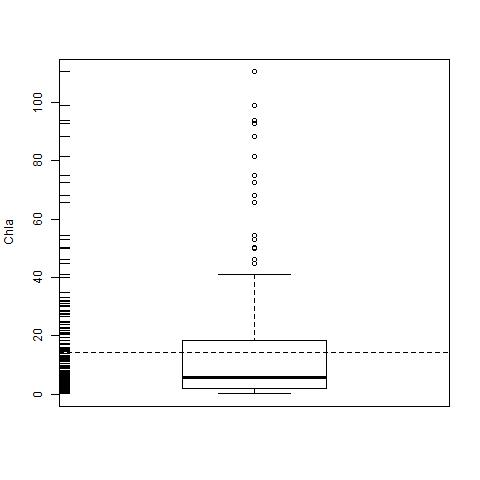
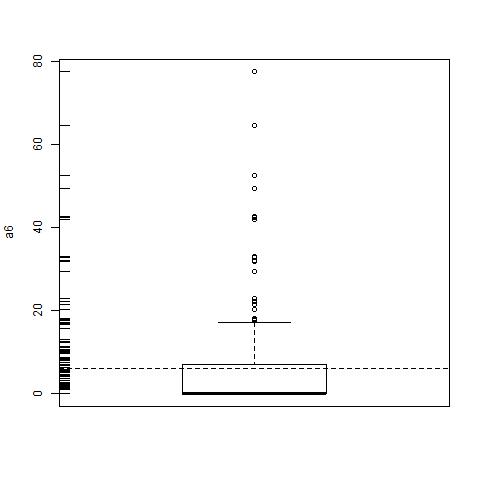
对于数值属性列，我们画出了所有的直方图和QQ图，保存在文件夹中，选取Chla 属性说明。

对于Chla属性，通过观察直方图和密度曲线，发现近似符合正态分布；使用QQ图可以验证这个结论。

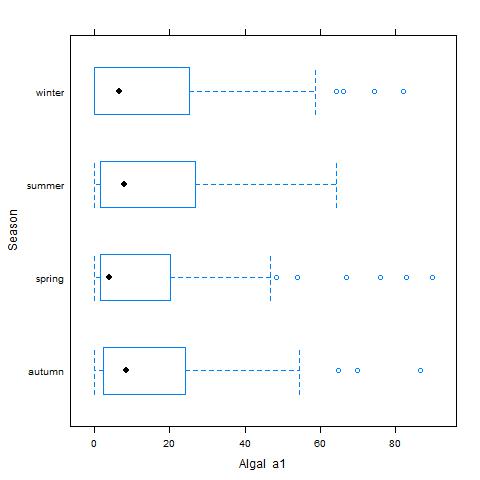
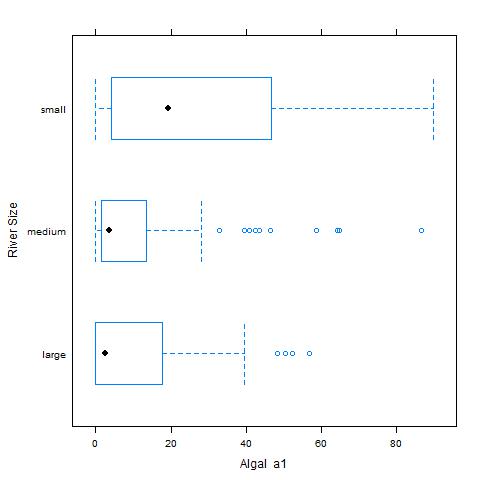
#### 画出数值属性的盒图

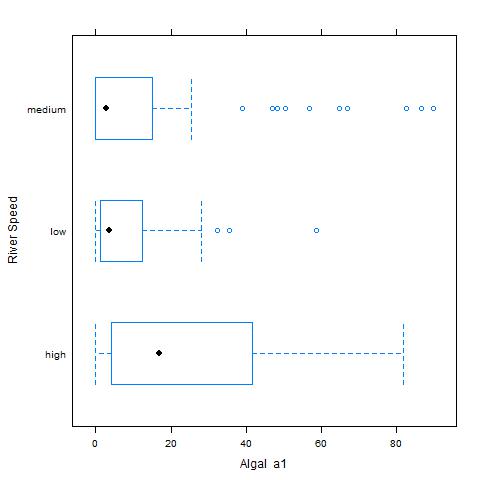
下面以chla 和a6对应的数据为例，其他属性的在文件夹中，他们都存在离群点，可以看出都是偏大的离群点。

#### 分别画出7种海藻和季节，河流大小，流速之间的条件盒图

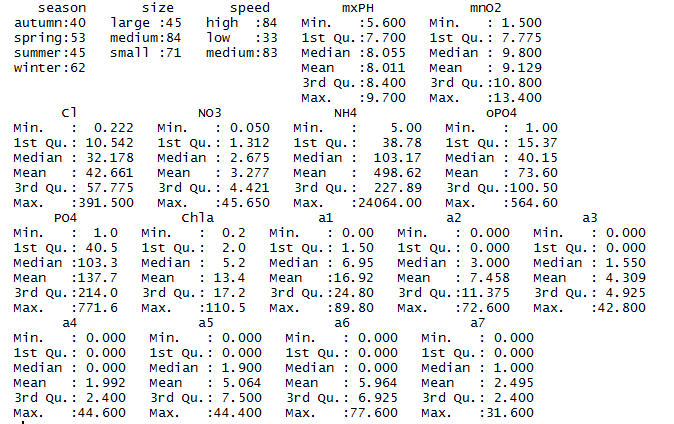
a 1 在夏天数量比较多，在小河流中比较多，在高流速的河流中数量多一些。



### (4) 通过数据对象之间的相似性来填补缺失值并可视化

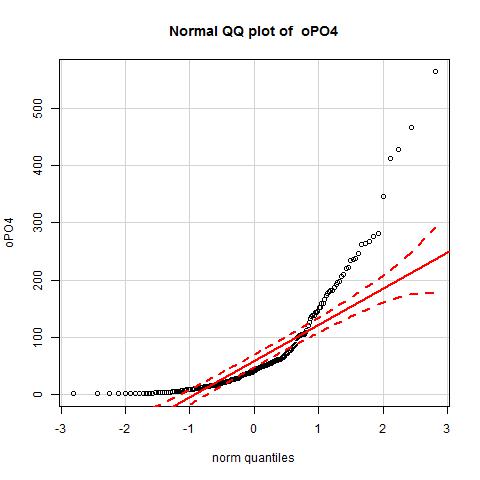
对于缺失样本我们采用使用数据对象之间的相似性来填补，处理结束后数据的摘要信息如下：



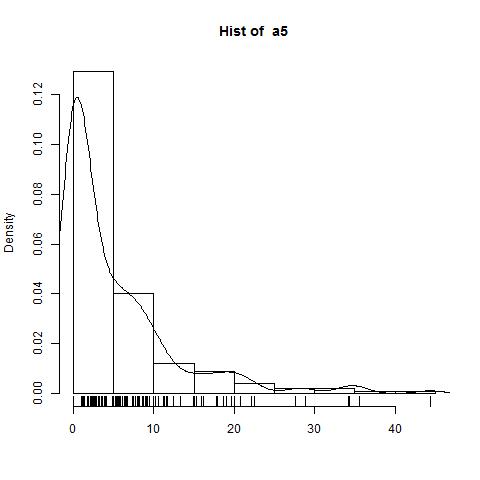
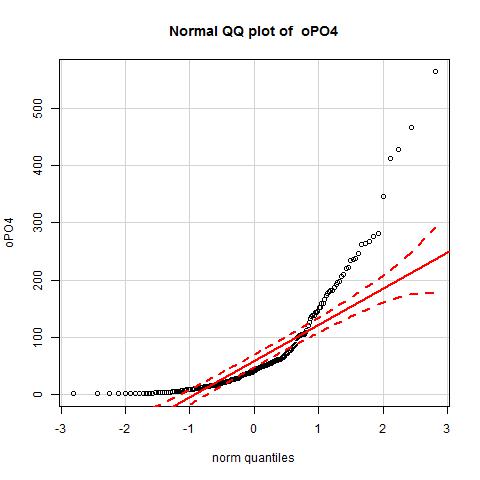
#### 使用直方图表示，使用QQ 图检验是否满足正态分布

对于数值属性列，我们画出了所有的直方图和Q-Q图，保存在文件夹中，下面我们选取其中两个说明。

对于oPO4 属性，通过观察直方图和密度曲线发现不符合正态分布，使用QQ图可以验证这个结论。

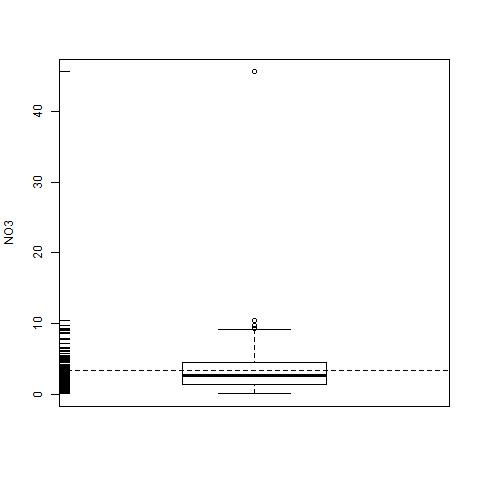
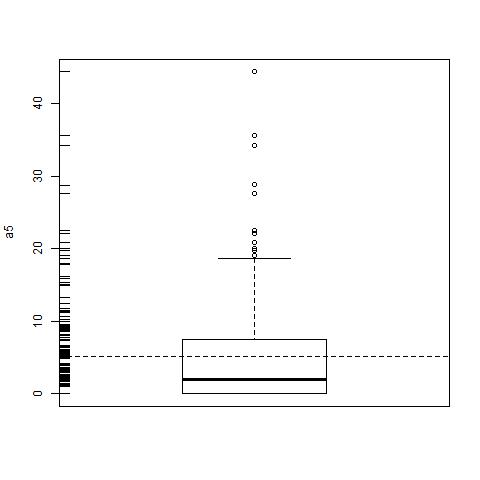


观察a5 属性，通过直方图我们可以看出来，a5的分布，不符合正态分布，其Q—Q 图也验证了不符合的结论。

#### 画出数值属性的盒图

下面以NO3 和a5对应的数据为例，其他属性的在文件夹中，他们都存在离群点，可以看出都是偏大的离群点。

#### 分别画出7种海藻和季节，河流大小，流速之间的条件盒图

a 3 在春天数量比较多，在中等河流中比较多，在低流速的河流中数量多一些。在中速和高速中分布差别不大。

