海藻数据的分析

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。

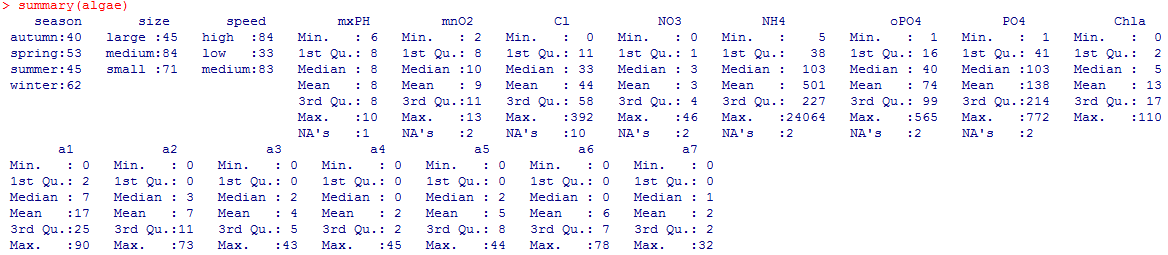
4. 提交内容

* 分析报告
* 分析程序
* 预处理后的数据集

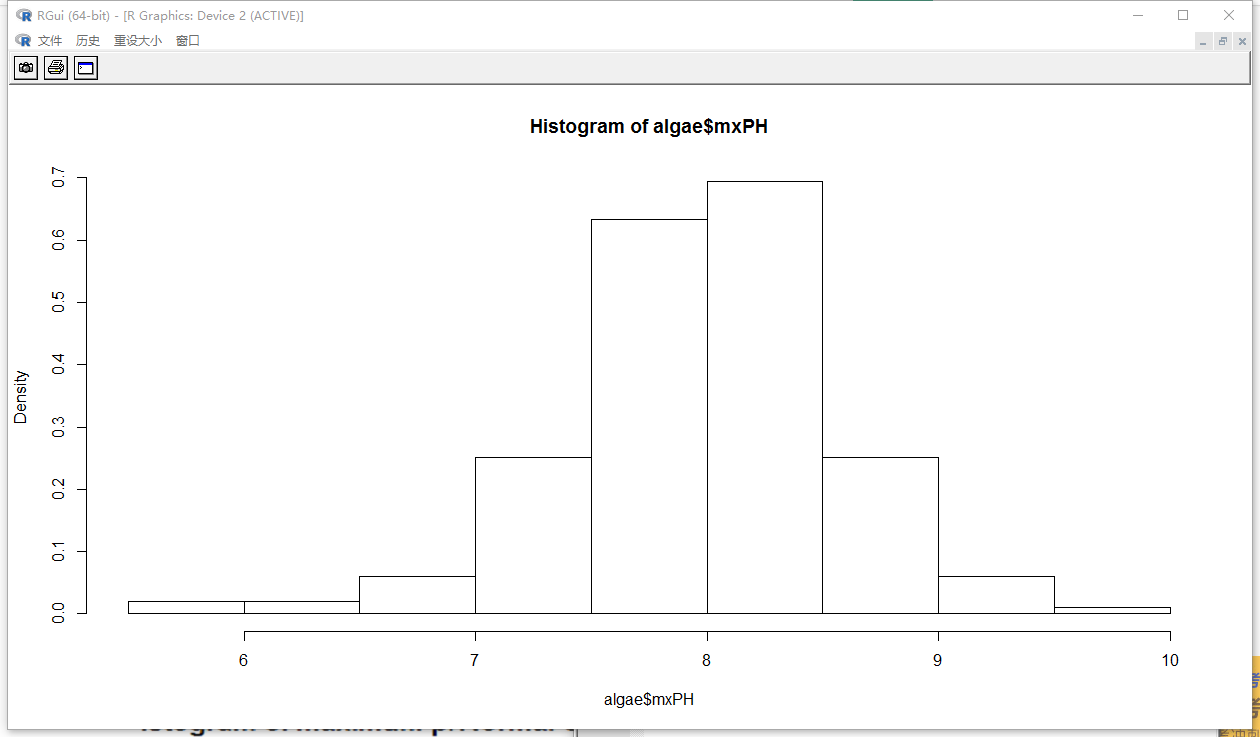
5.实验过程

1.加载所需库文件：

2.读取实验数据并获取数据摘要：

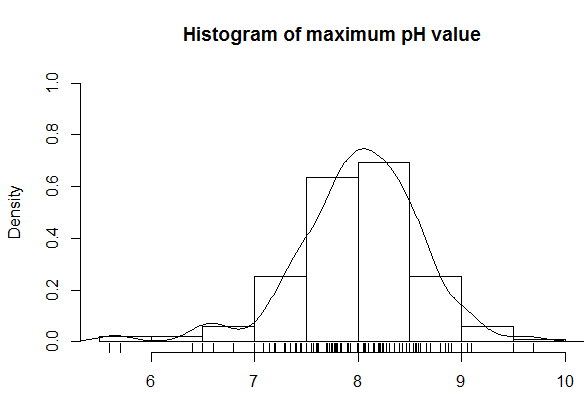


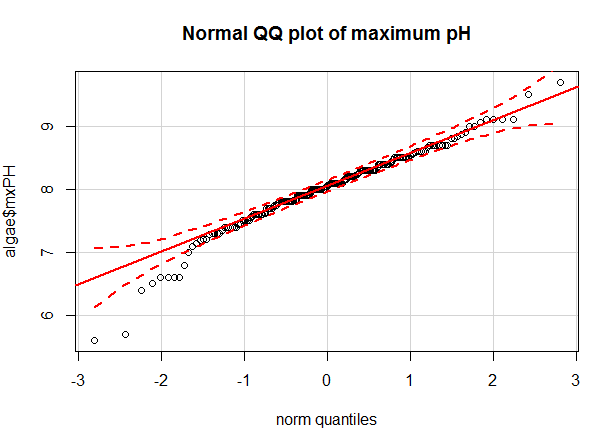
3.绘制mxPH的直方图：



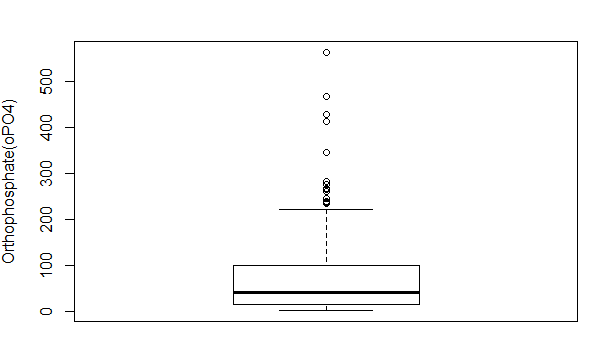
4.绘制QQ图：



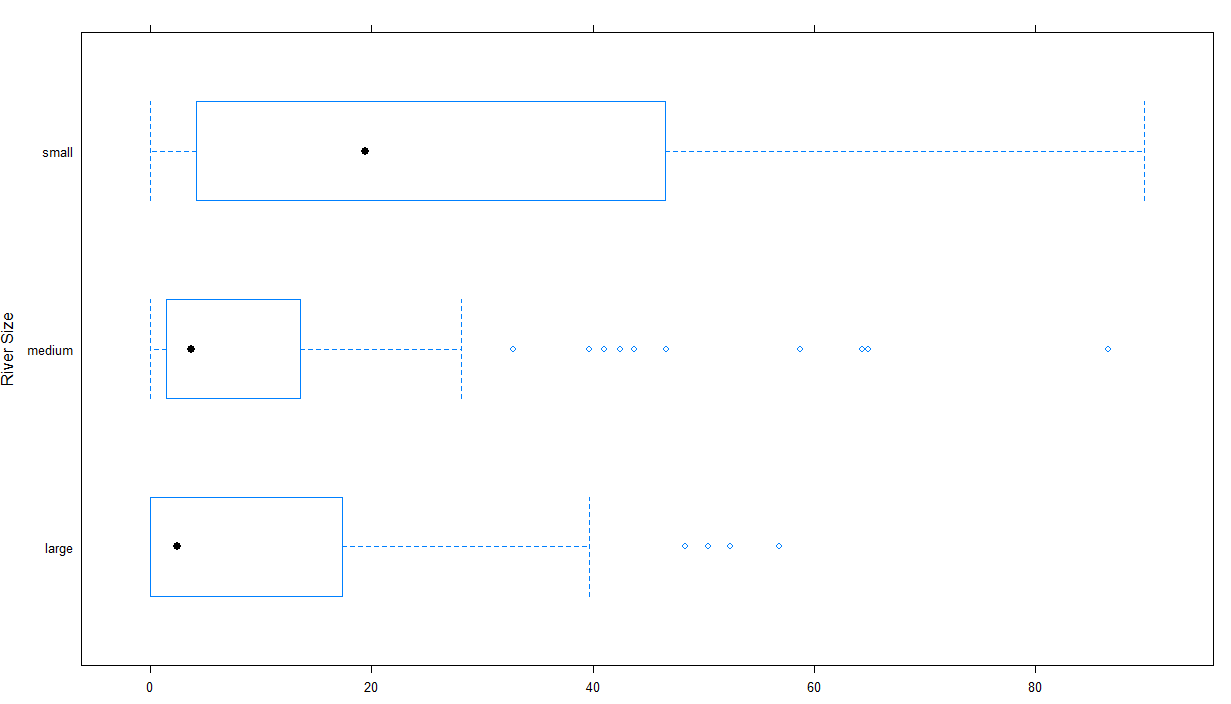




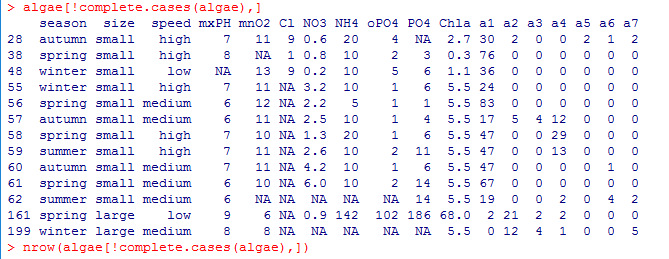
5.绘制OPo4的盒图：

6.绘制七种海藻和size之间的条件盒图

7.将缺失值剔除

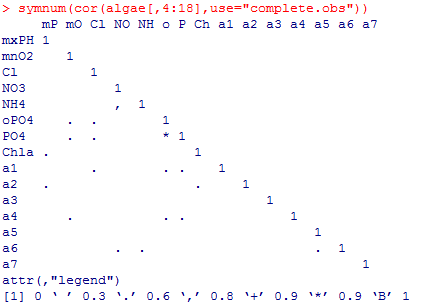
 

8.用高频率来填补缺失值

9.通过属性的相关关系填补缺失值



相关值矩阵：



10.通过案例之间的相似关系填补缺失值

