海藻实验报告

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

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。

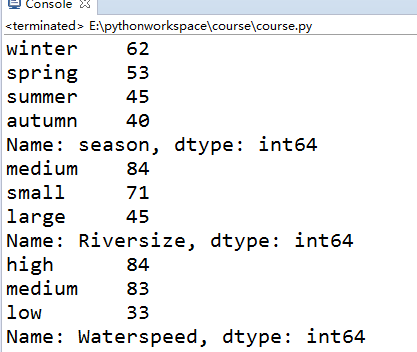
4.实验报告

4.1准备工作

将txt文件中的数据转换为csv文件中的数据，并将XXXXXX设为空值。

4.2数据分析

标称属性：

标称属性为：season( winter、spring、autumn、summer)

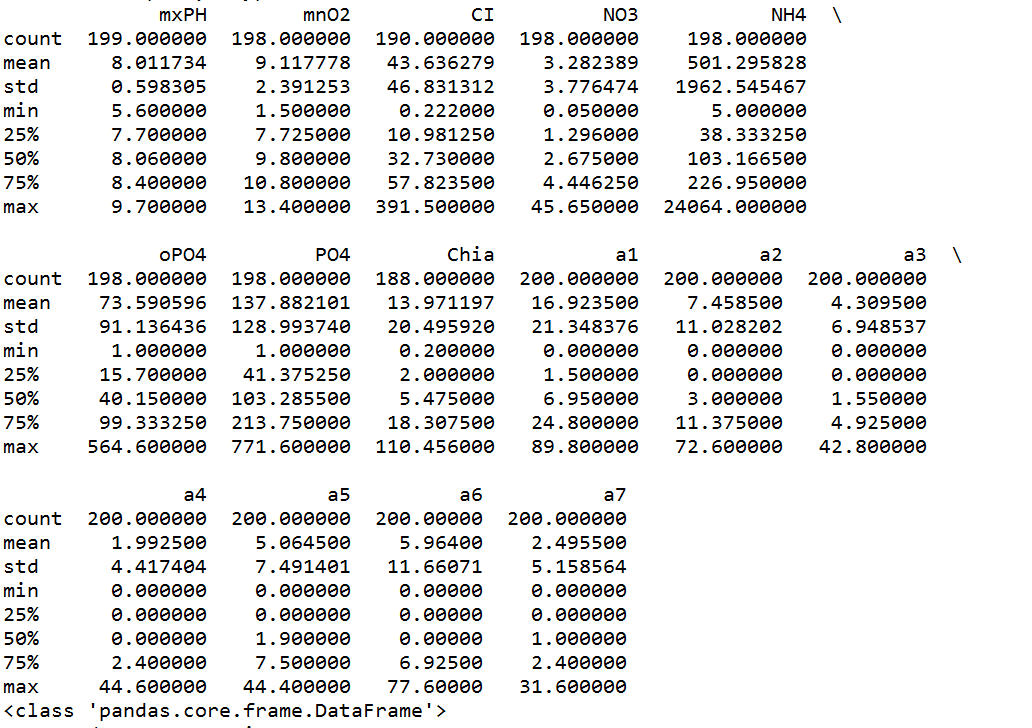
Riversize(small、medium、large)；

Waterspeed(low、medium、high)

然后统计每个标称属性每个取值的数据量。

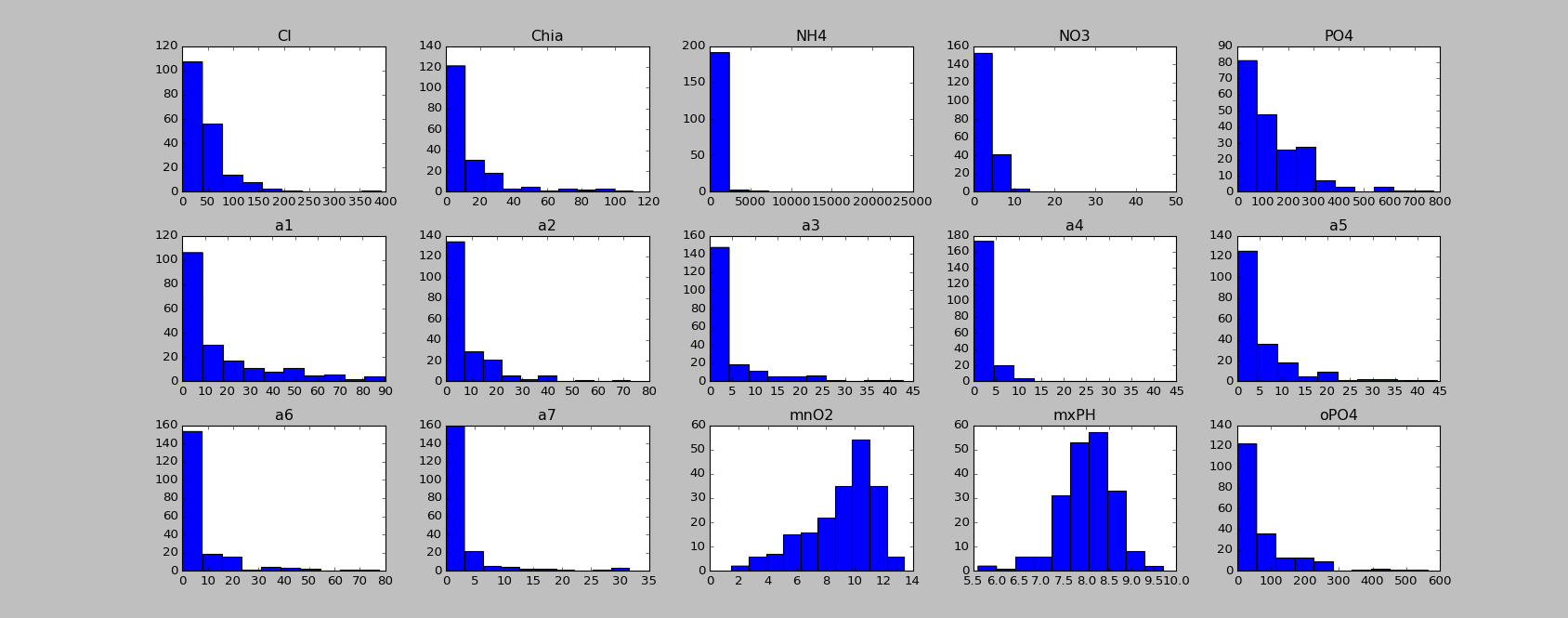
数值属性：mxPH、mnO2、CI、NO3、NH4、oPO4、PO4、Chia、a1、a2、a3、a4、a5、a6、a7.

统计数值属性的最大、最小、均值、中位数、四分位数及缺失值的个数。

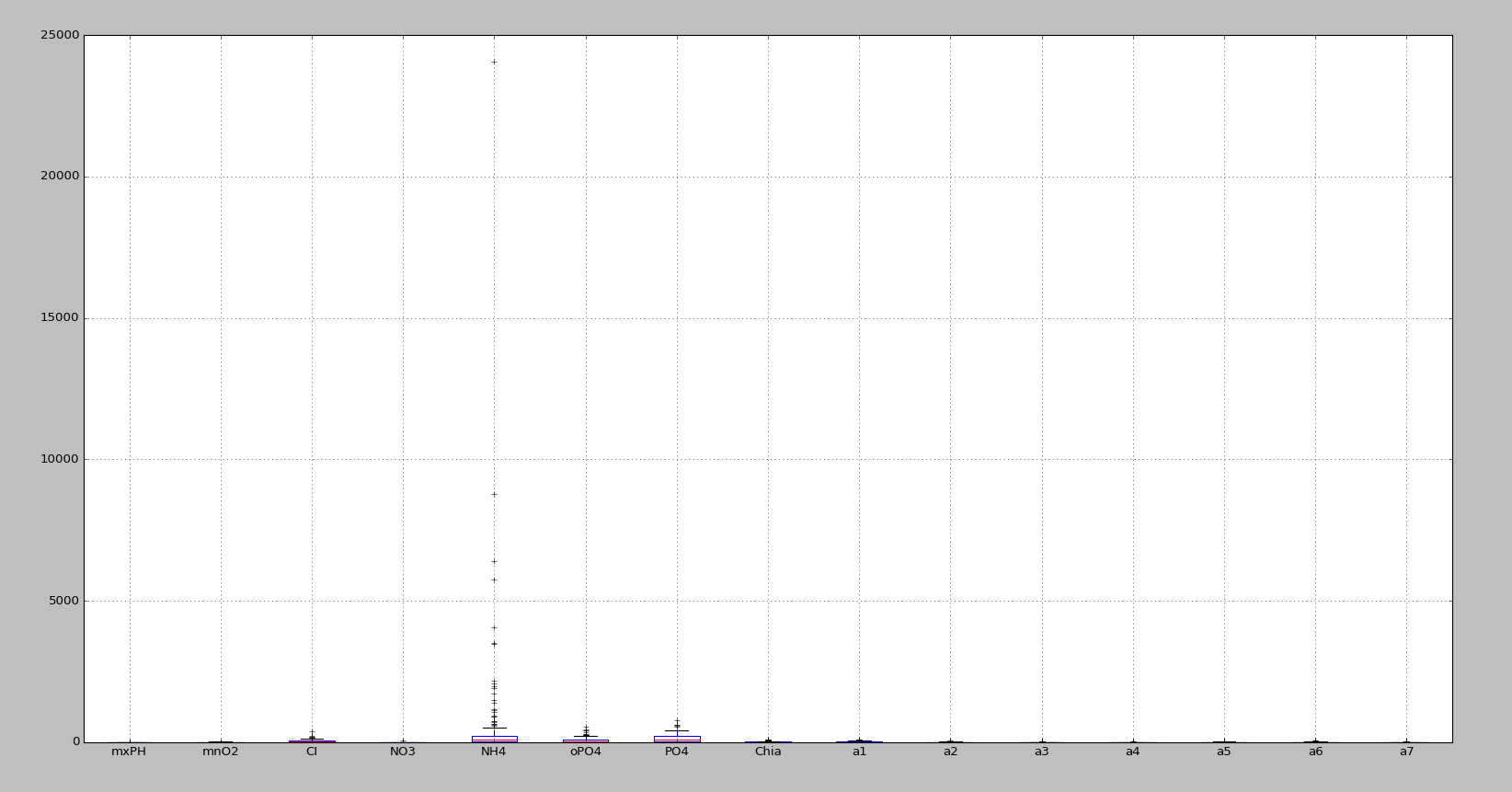


数据可视化：

对于数值属性，绘制直方图和盒图。

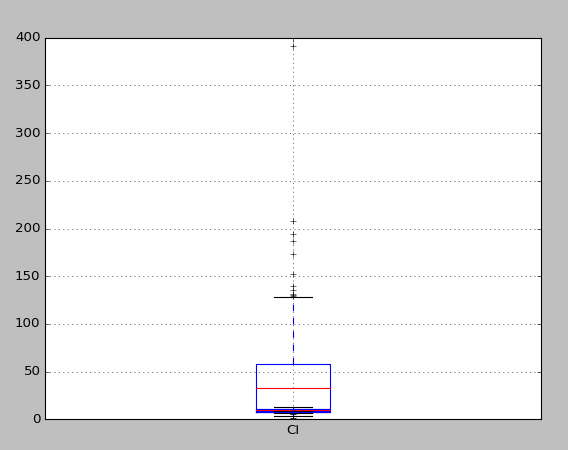
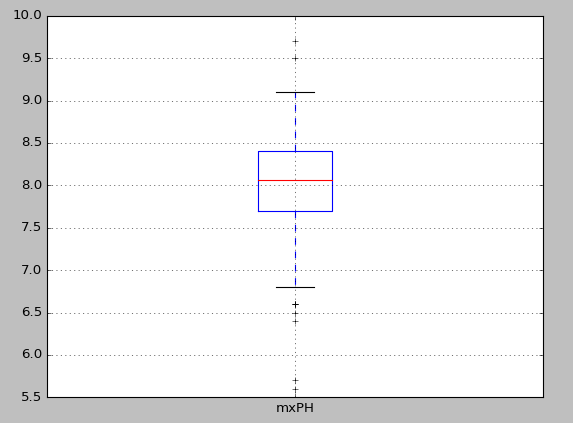
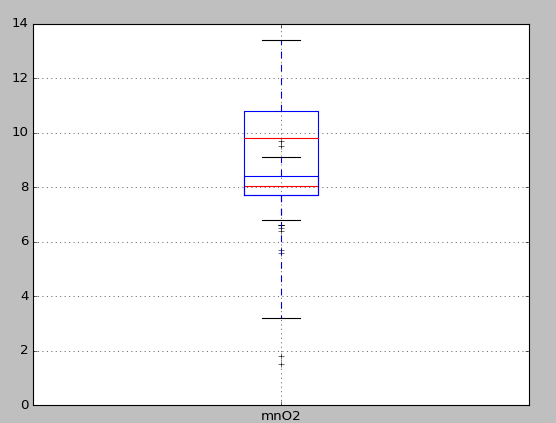


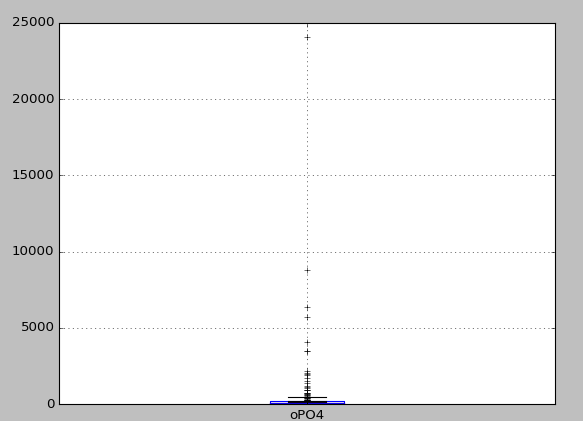
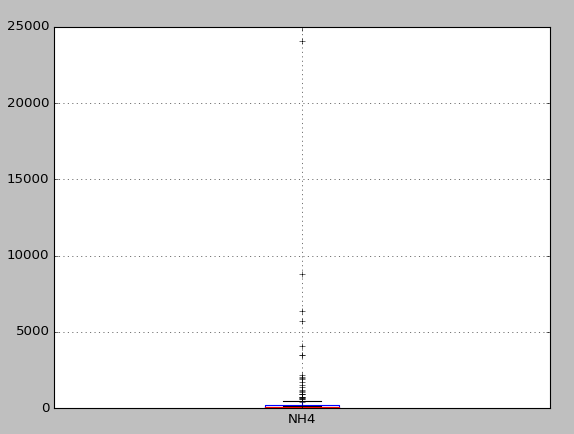
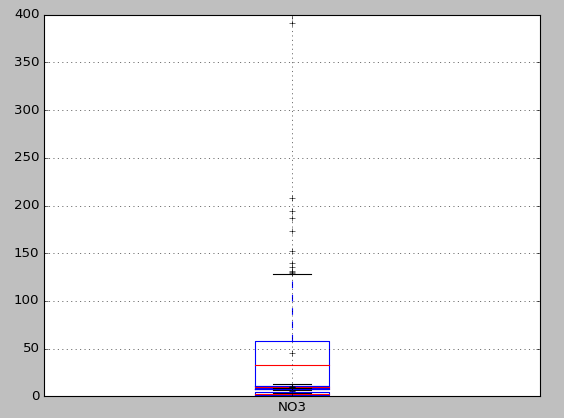
直方图

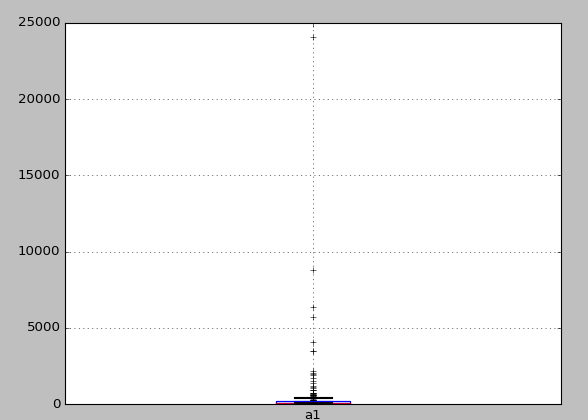
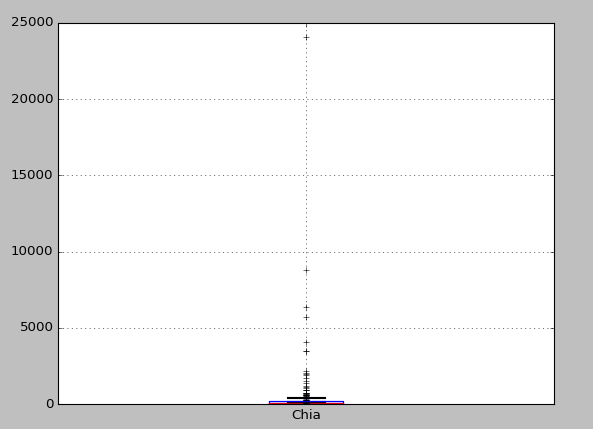
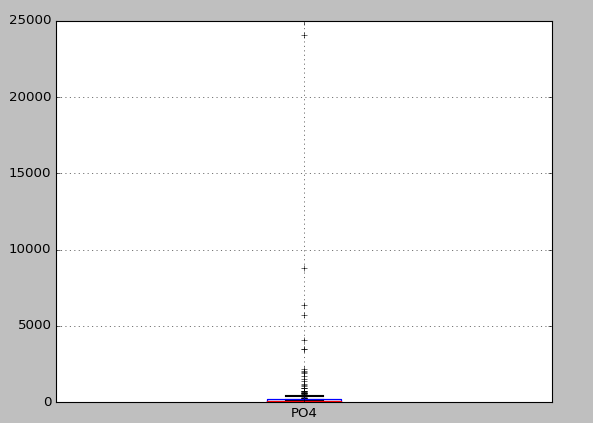


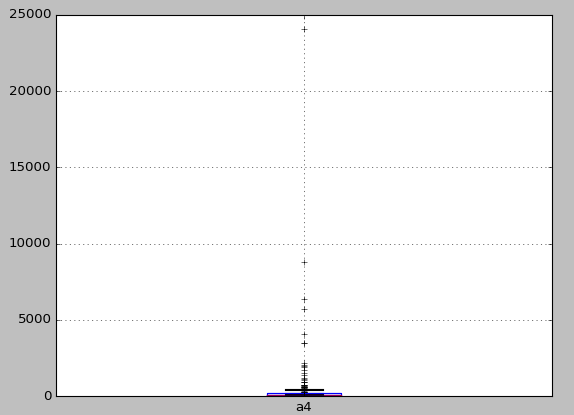
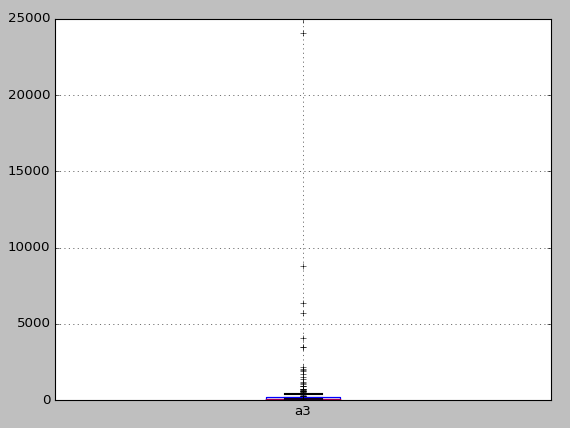
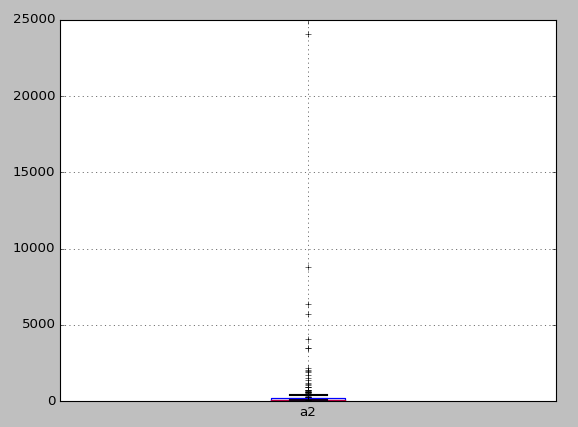
盒图

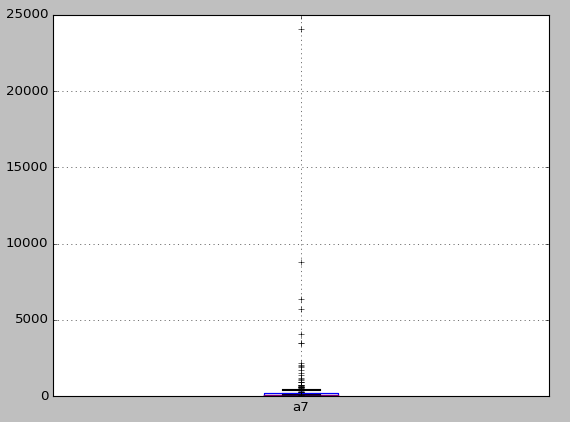
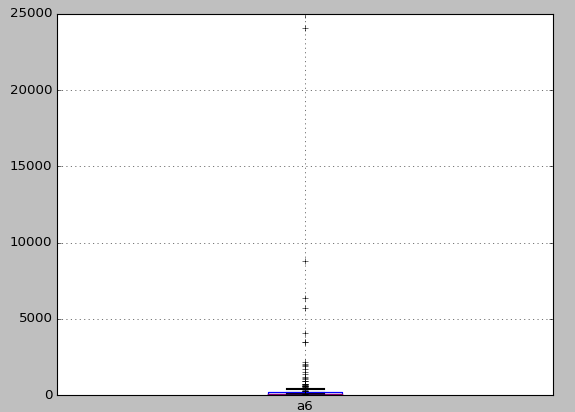
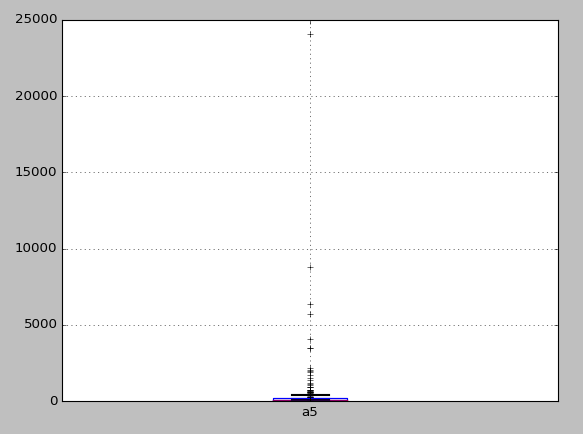
由于离群点的存在，整体图效果不清楚，每个属性的盒图如下：





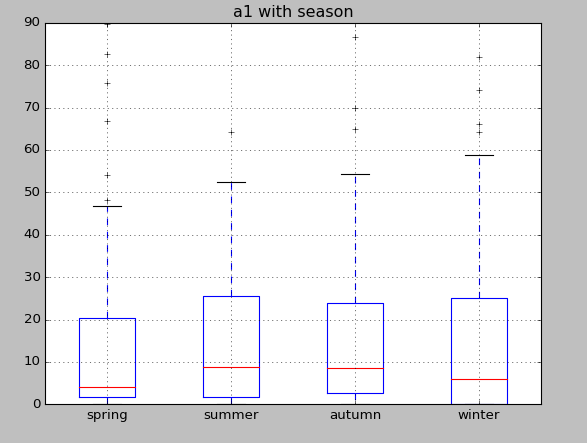


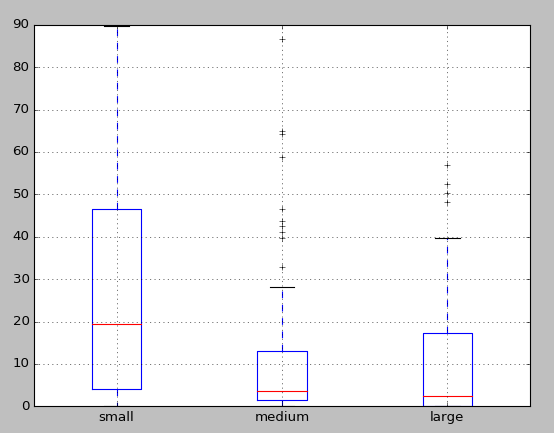


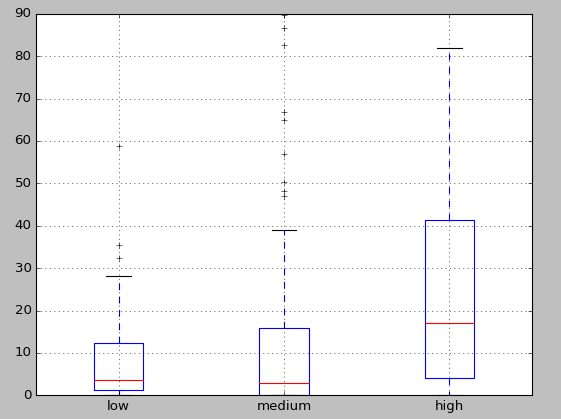


条件盒图：

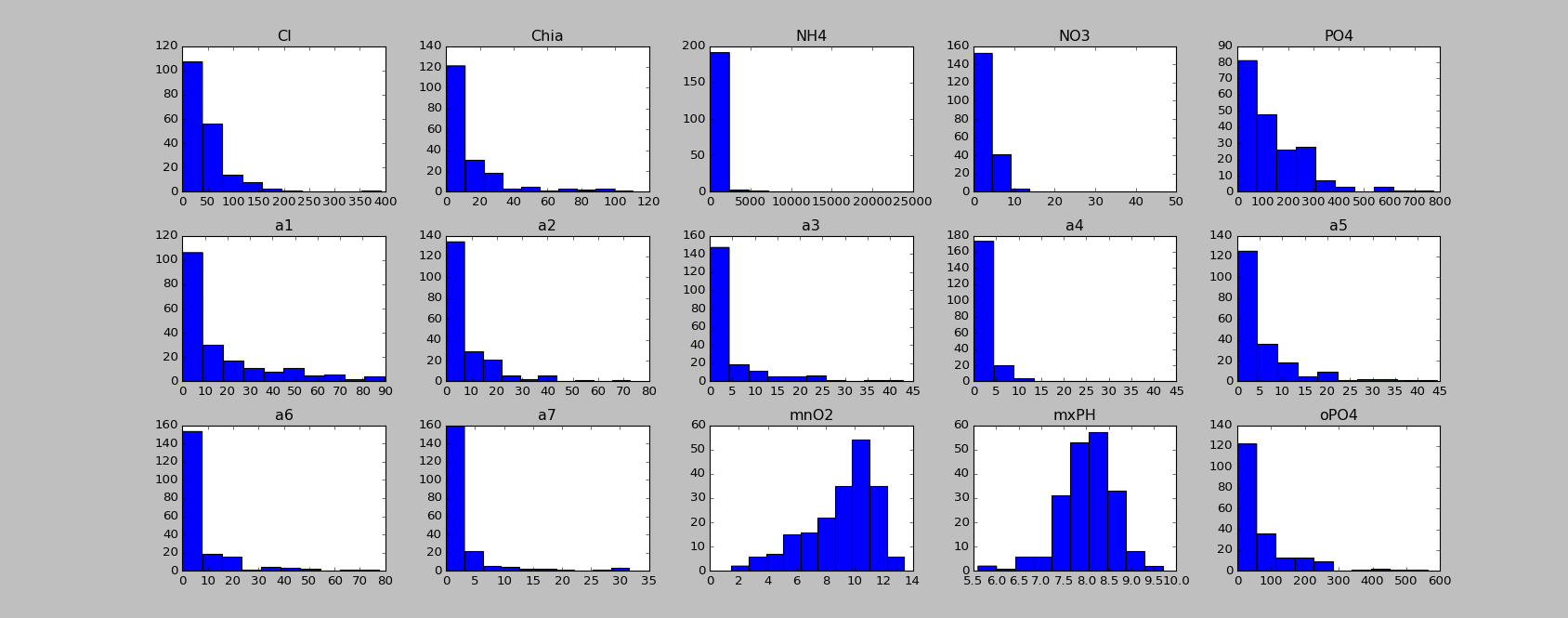
对于属性a1，分别画出它与season,riversize,waterspeed的条件盒图，a2-a7类似。



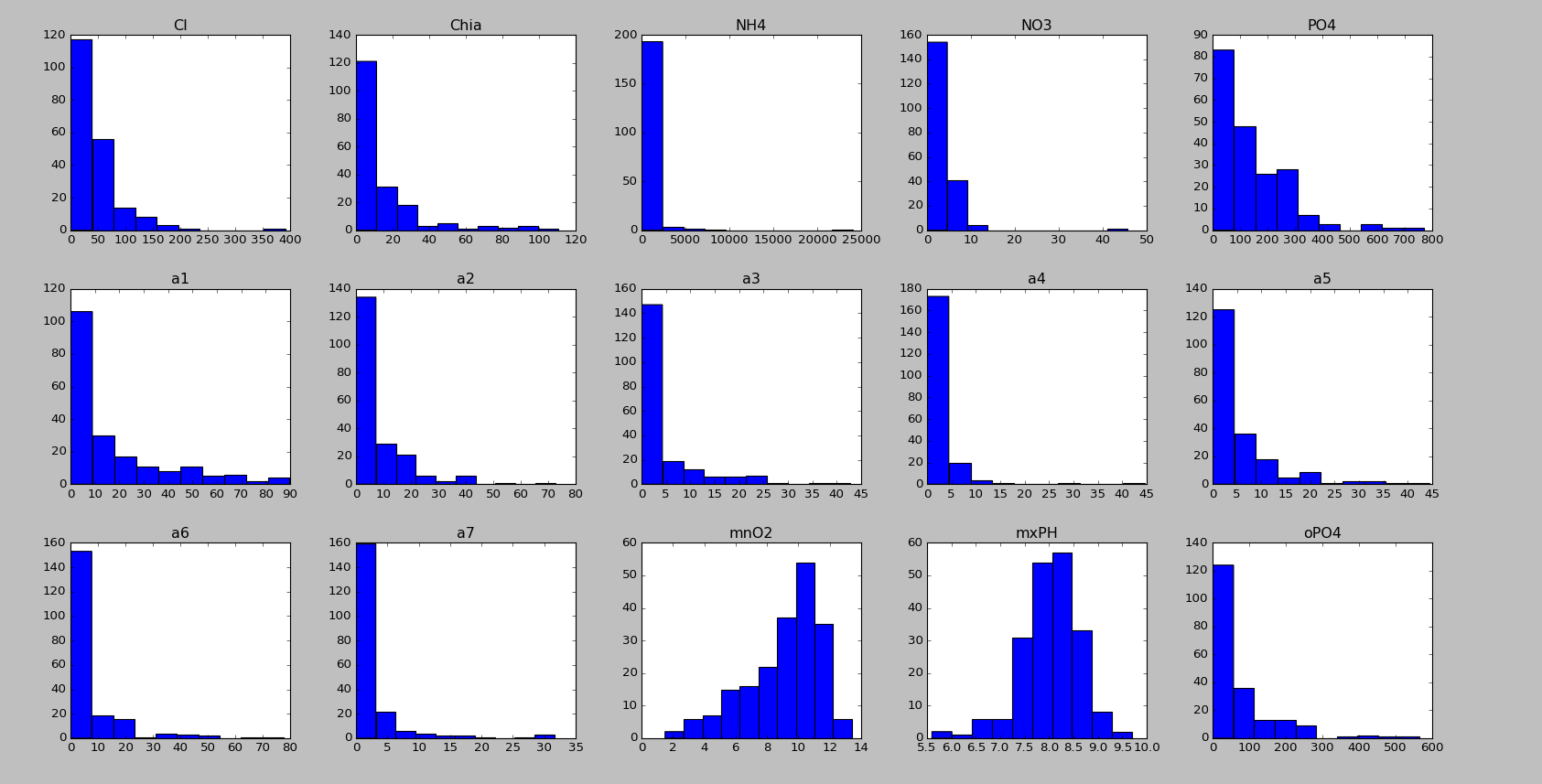




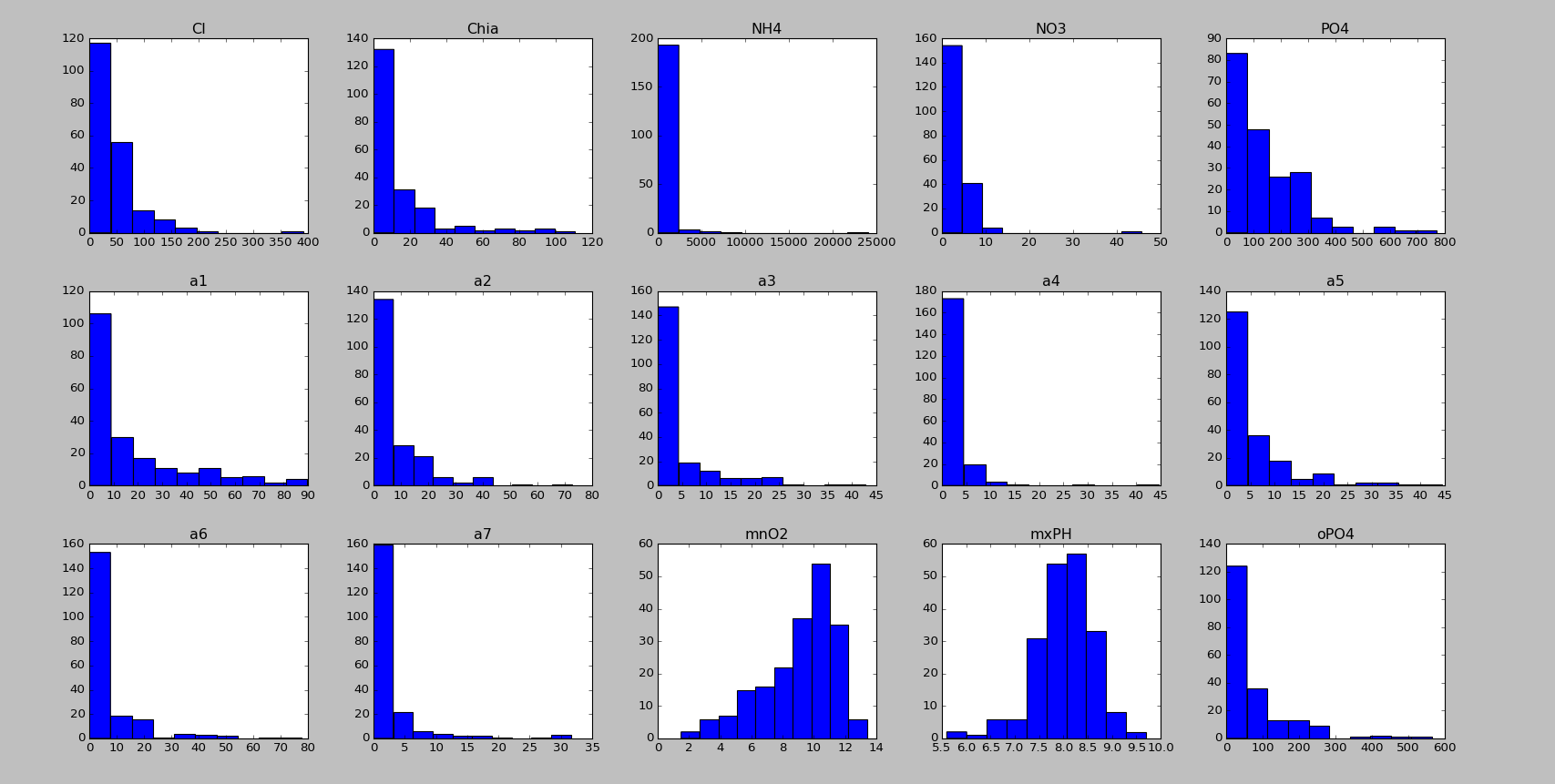
将属性值为空的数据删除，数据直方图可视化：



用众数填补空缺值，可视化直方图：

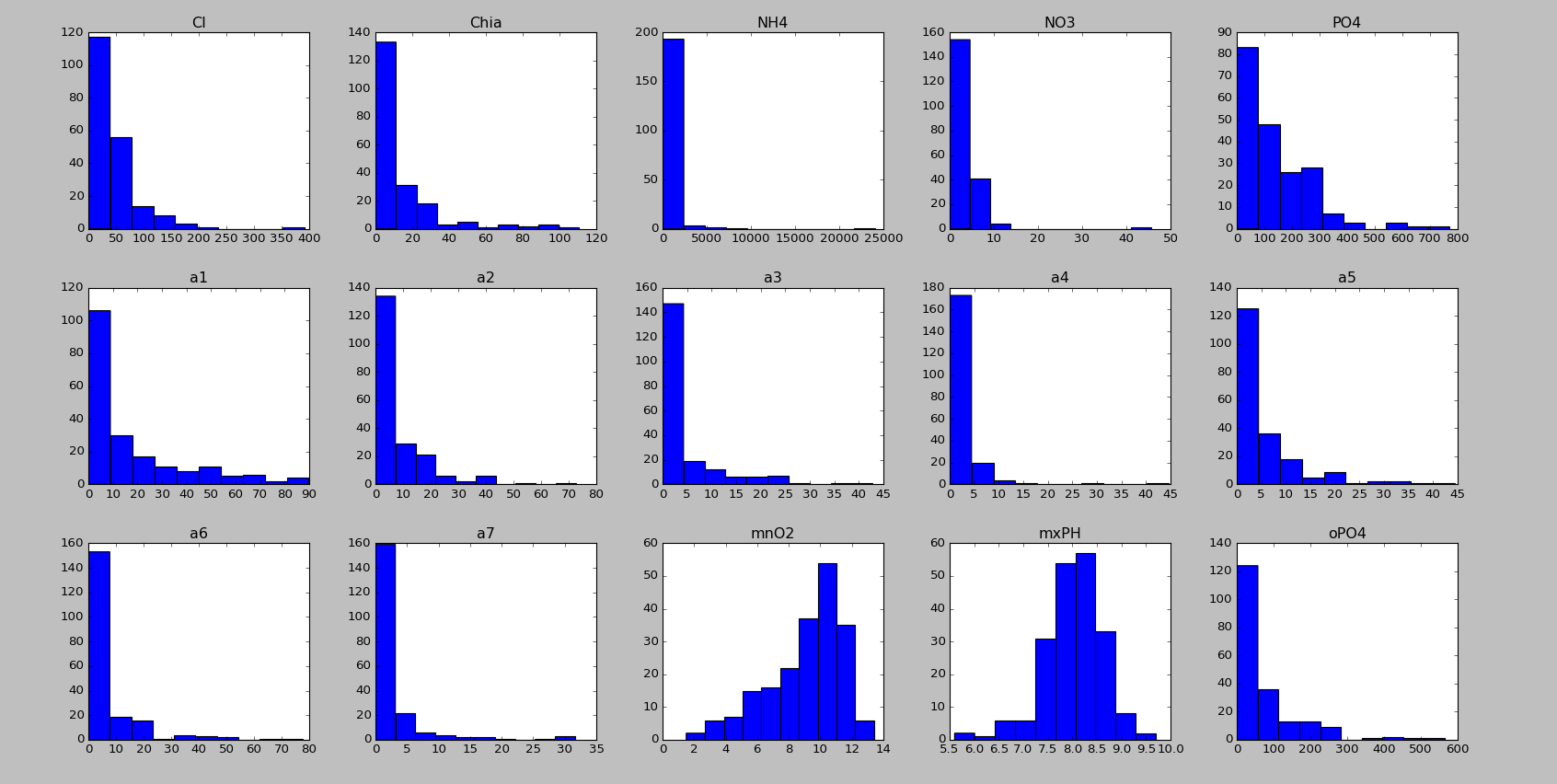


通过属性相关关系来填补缺失值，比如用随机森林填补Chia的值，数据直方图可视化：



通过数据对象之间的相似性来填补缺失值。

通过计算空缺数据的mxPH,mnO2,CI,NO3,NH4,oPO4,PO4和完整数据的对应属性之间的欧式距离，选择和完整数据对应属性的距离最小的那条数据的对应的Chia的值，将对应的空缺值填补，数据直方图如下：



5.实验结论

本次实验，主要是对数据缺失值的处理，缺失值的处理的有很多种方法，具体到使用哪一种方法，还要根据数据的属性，假如是标称属性，那么可以采用众数填补，对于数值属性，可以用随机森林来进行填补，通过其他属性进行构建随机森林，来对缺失值进行预测，我觉得这是最好的进行缺失值处理的方法。当然假如数据量非常大时，而缺失值非常少的话，那么可以直接将缺失的数据删除，另外，在实际的数据挖掘中，假如缺失值对应的那个属性，对于数据挖掘的任务不是很重要，那么可以直接去掉这个属性，不把它作为数据挖掘的特征。只有将数据处理好，构建出好的特征，这才是数据挖掘中最重要的一点。