实验数据报告

姓名：王旗林

学号：2120151041

邮箱：qilin\_bit@126.com

**一、实验内容**

**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**

* 1. **数据可视化和摘要**

1）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。

2）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。

**二、实验环境**

Python版本：Python3.5.1

开发环境：Windows配置Eclipse+Pydev

**三、实验结果**

1. **数据摘要**

**对于标称属性：**

季节频数：

winter 62

spring 53

summer 45

autumn 40

dtype: int64

河流大小频数：

medium 84

small 71

large 45

dtype: int64

河水速度频数：

high 84

medium 83

low 33

dtype: int64

**对于数值属性：**

mxPH mnO2 Cl NO3 NH4 \

count 199.000000 198.000000 190.000000 198.000000 198.000000

mean 8.011734 9.117778 43.636279 3.282389 501.295828

std 0.598305 2.391253 46.831311 3.776474 1962.545461

min 5.600000 1.500000 0.222000 0.050000 5.000000

25% NaN NaN NaN NaN NaN

50% NaN NaN NaN NaN NaN

75% NaN NaN NaN NaN NaN

max 9.700000 13.400000 391.500000 45.650000 24064.000000

oPO4 PO4 Chla a1 a2 a3 \

count 198.000000 198.000000 188.000000 200.000000 200.000000 200.000000

mean 73.590596 137.882101 13.971197 16.923500 7.458500 4.309500

std 91.136436 128.993738 20.495920 21.348376 11.028202 6.948537

min 1.000000 1.000000 0.200000 0.000000 0.000000 0.000000

25% NaN NaN NaN 1.500000 0.000000 0.000000

50% NaN NaN NaN 6.950000 3.000000 1.550000

75% NaN NaN NaN 24.800000 11.375000 4.925000

max 564.599980 771.599980 110.456000 89.800000 72.600000 42.800000

a4 a5 a6 a7

count 200.000000 200.000000 200.00000 200.000000

mean 1.992500 5.064500 5.96400 2.495500

std 4.417404 7.491401 11.66071 5.158564

min 0.000000 0.000000 0.00000 0.000000

25% 0.000000 0.000000 0.00000 0.000000

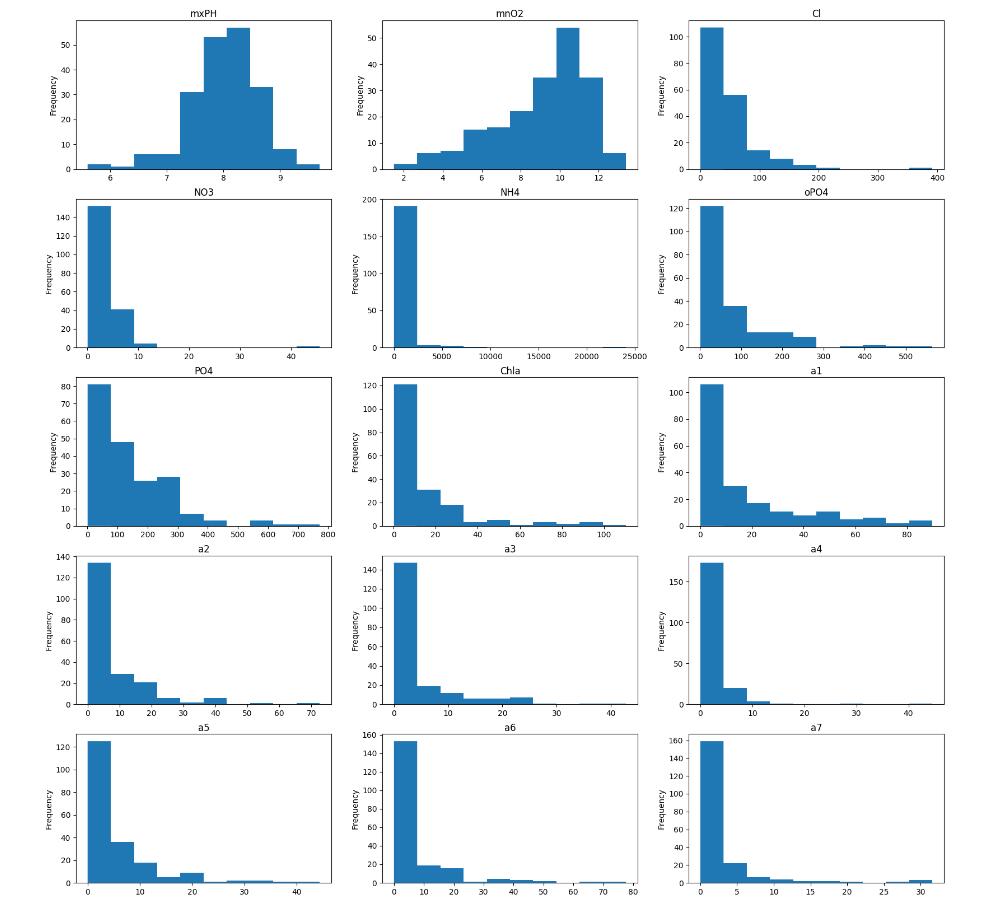
50% 0.000000 1.900000 0.00000 1.000000

75% 2.400000 7.500000 6.92500 2.400000

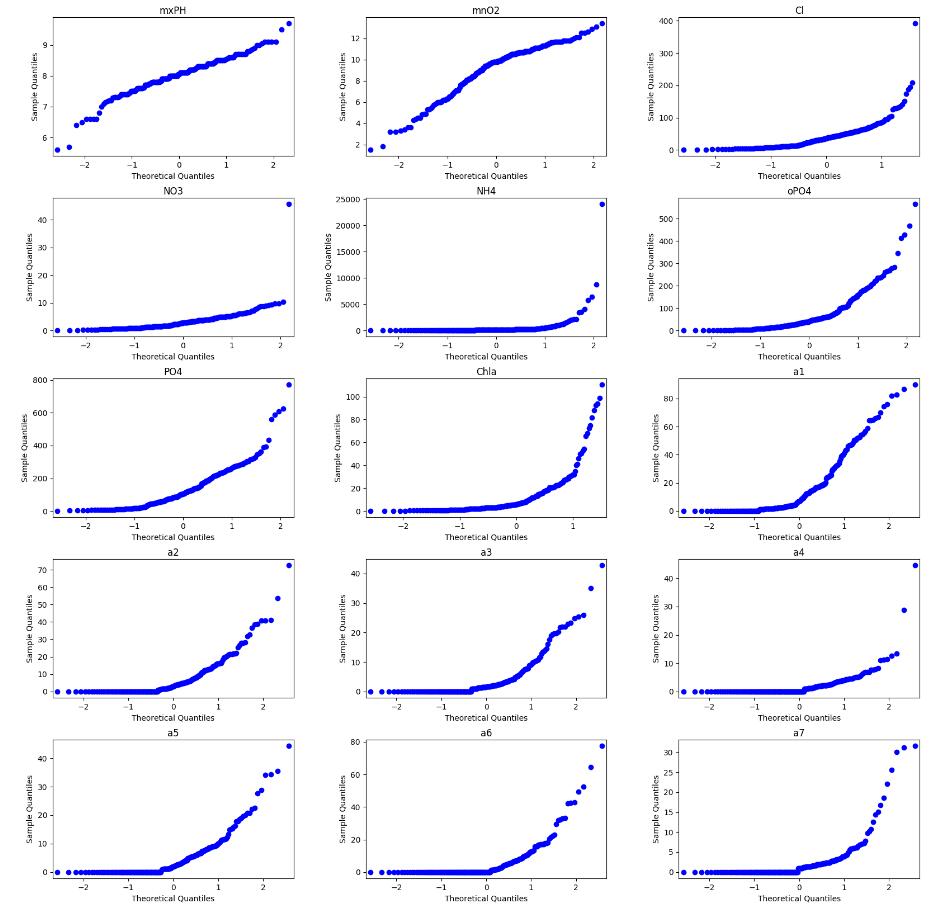
max 44.600000 44.400000 77.60000 31.600000

1. **数据可视化**

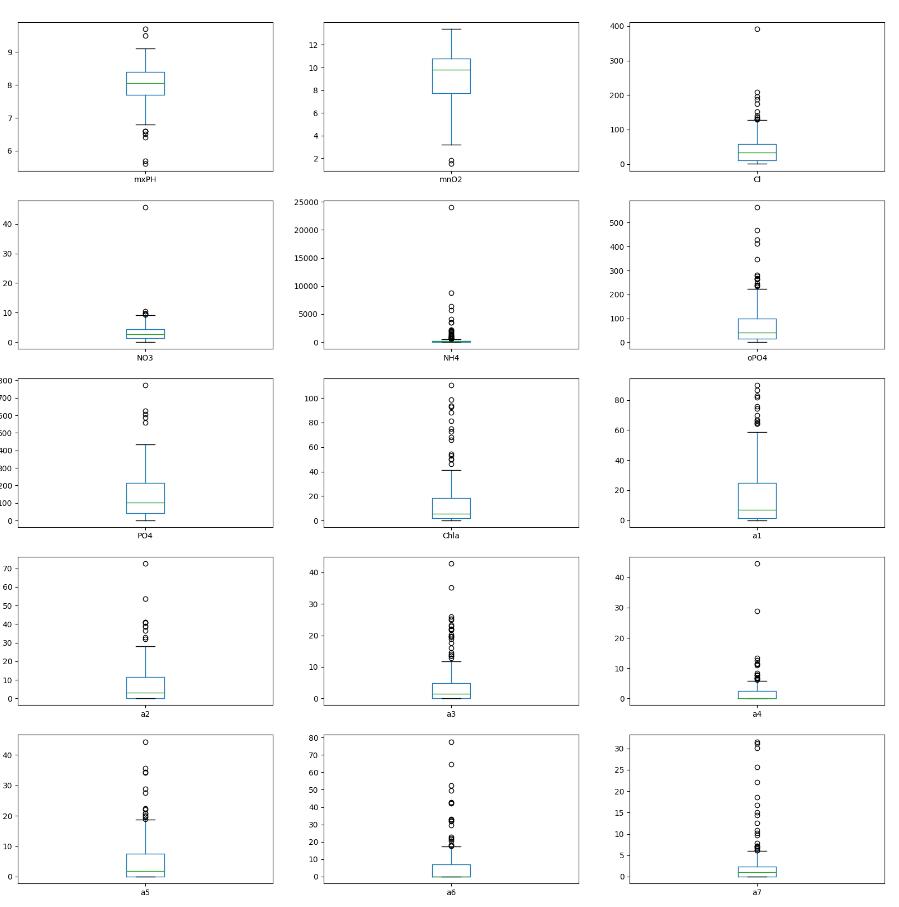
**数据直方图**

****

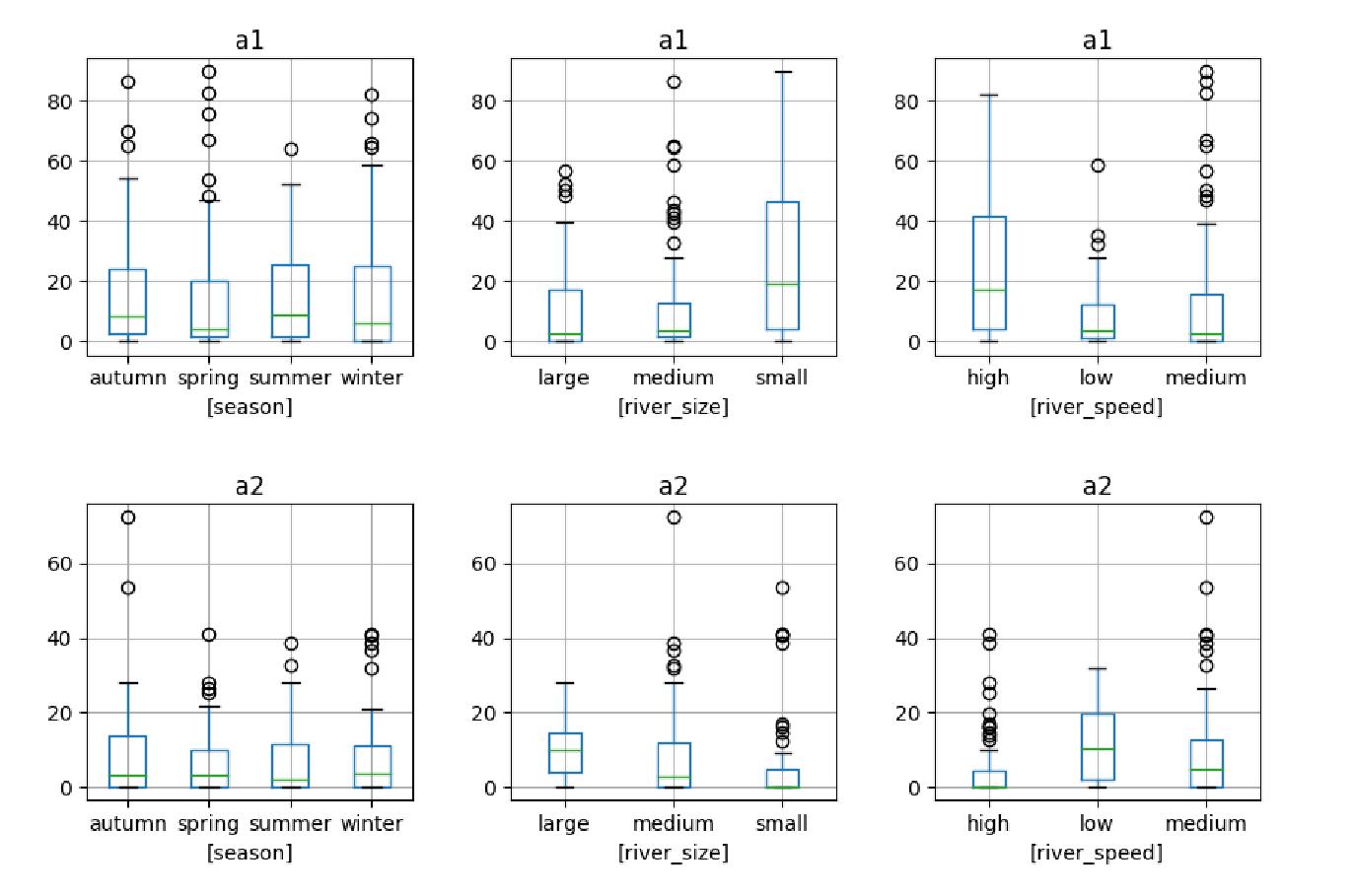
**数据qq图：**

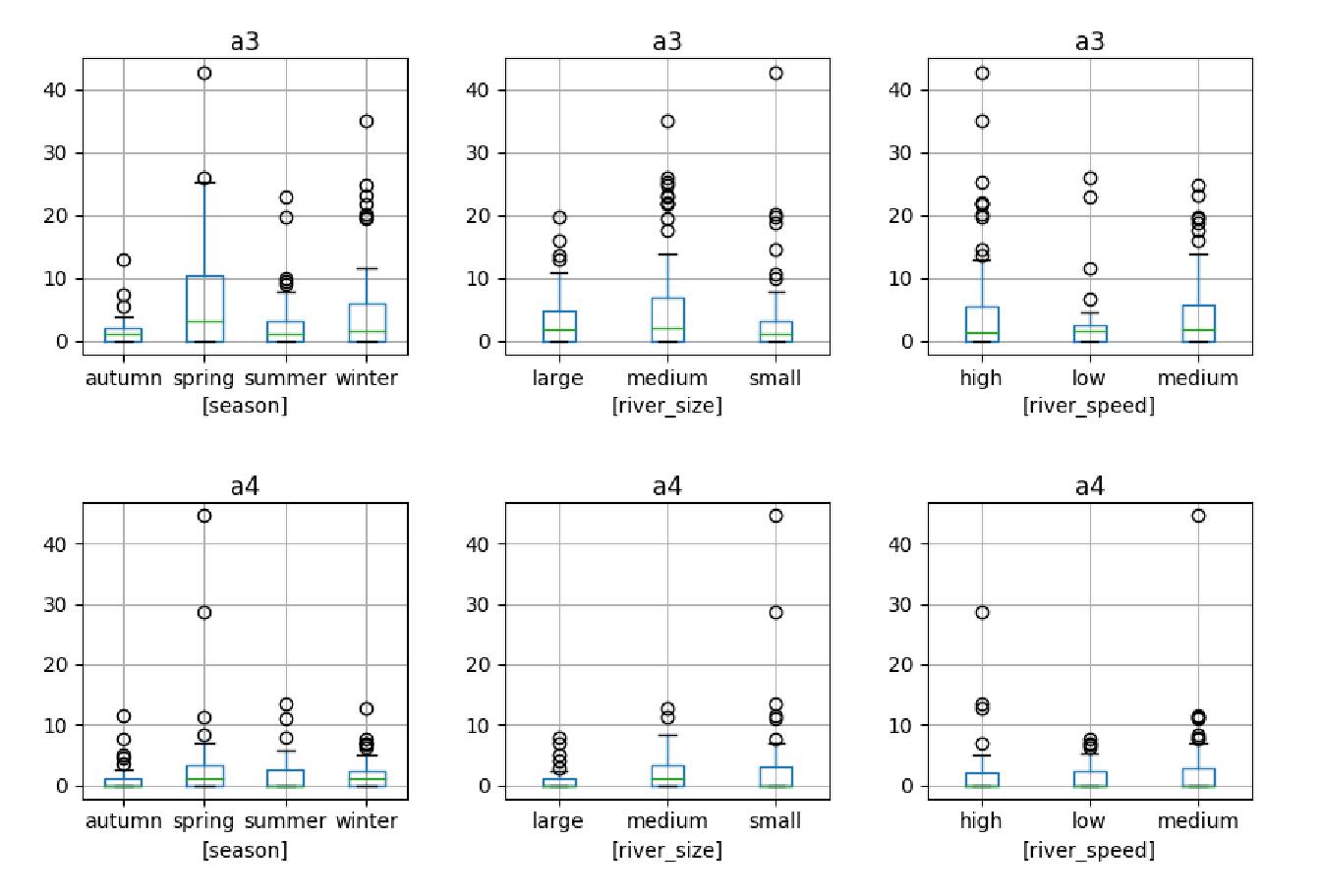
****

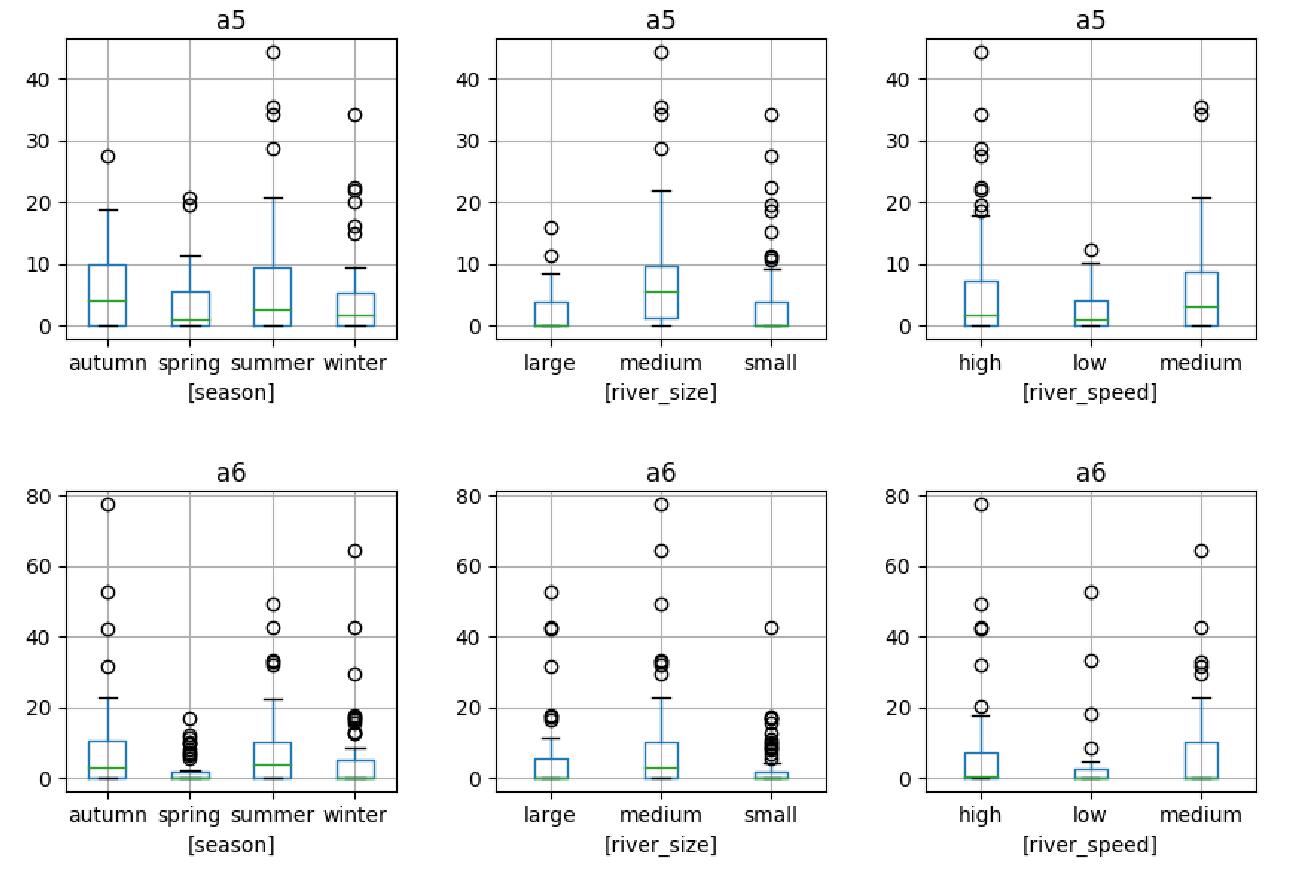
**数据盒图**

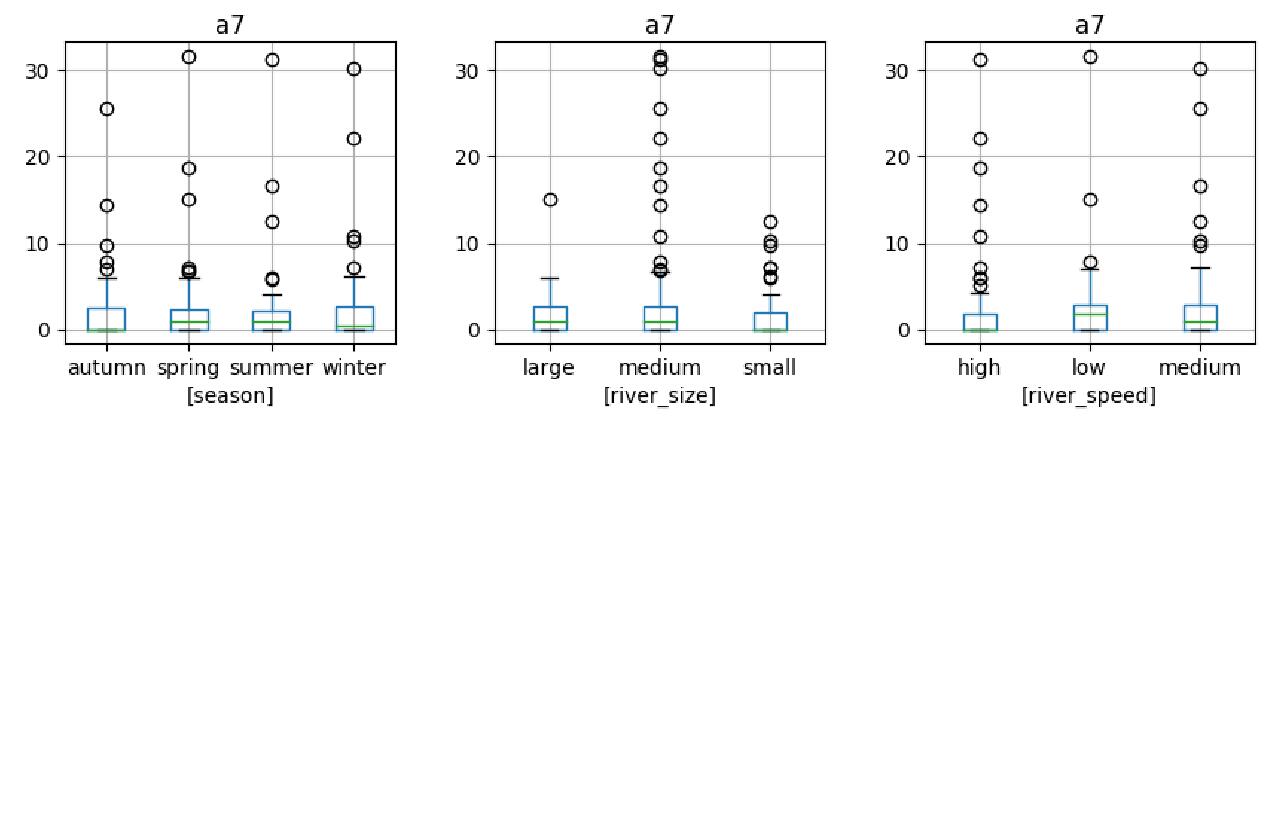
****

**条件盒图**

****

****

****

****

**3.缺失值处理**

**3.1 将缺失部分剔除**

结果：delete.csv

**3.2 用最高频率值来填补缺失值**

结果：most.csv

**3.3 通过属性的相关关系来填补缺失值**

结果：corelation.csv

**3.4 通过数据对象之间的相似性来填补缺失值**

结果：similarity.csv