###仓库说明：

`algae.R`是对原始数据处理代码，并对缺失的数据进行4种填充，\*\*填充完缺失的数据在\\*.csv文件中，文件名中指出了处理缺失数据的方法\*\*。

`algae\_fixed.R`是对填充完的数据进行可视化处理代码。

`algae doc.docx`是总结文档。

1. 数据摘要

## season size speed mxPH mnO2

## autumn:40 large :45 high :84 Min. :5.600 Min. : 1.500

## spring:53 medium:84 low :33 1st Qu.:7.700 1st Qu.: 7.725

## summer:45 small :71 medium:83 Median :8.060 Median : 9.800

## winter:62 Mean :8.012 Mean : 9.118

## 3rd Qu.:8.400 3rd Qu.:10.800

## Max. :9.700 Max. :13.400

## NA's :1 NA's :2

## Cl NO3 NH4 oPO4

## Min. : 0.222 Min. : 0.050 Min. : 5.00 Min. : 1.00

## 1st Qu.: 10.981 1st Qu.: 1.296 1st Qu.: 38.33 1st Qu.: 15.70

## Median : 32.730 Median : 2.675 Median : 103.17 Median : 40.15

## Mean : 43.636 Mean : 3.282 Mean : 501.30 Mean : 73.59

## 3rd Qu.: 57.824 3rd Qu.: 4.446 3rd Qu.: 226.95 3rd Qu.: 99.33

## Max. :391.500 Max. :45.650 Max. :24064.00 Max. :564.60

## NA's :10 NA's :2 NA's :2 NA's :2

## PO4 Chla a1 a2

## Min. : 1.00 Min. : 0.200 Min. : 0.00 Min. : 0.000

## 1st Qu.: 41.38 1st Qu.: 2.000 1st Qu.: 1.50 1st Qu.: 0.000

## Median :103.29 Median : 5.475 Median : 6.95 Median : 3.000

## Mean :137.88 Mean : 13.971 Mean :16.92 Mean : 7.458

## 3rd Qu.:213.75 3rd Qu.: 18.308 3rd Qu.:24.80 3rd Qu.:11.375

## Max. :771.60 Max. :110.456 Max. :89.80 Max. :72.600

## NA's :2 NA's :12

## a3 a4 a5 a6

## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.000

## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000

## Median : 1.550 Median : 0.000 Median : 1.900 Median : 0.000

## Mean : 4.309 Mean : 1.992 Mean : 5.064 Mean : 5.964

## 3rd Qu.: 4.925 3rd Qu.: 2.400 3rd Qu.: 7.500 3rd Qu.: 6.925

## Max. :42.800 Max. :44.600 Max. :44.400 Max. :77.600

##

## a7

## Min. : 0.000

## 1st Qu.: 0.000

## Median : 1.000

## Mean : 2.495

## 3rd Qu.: 2.400

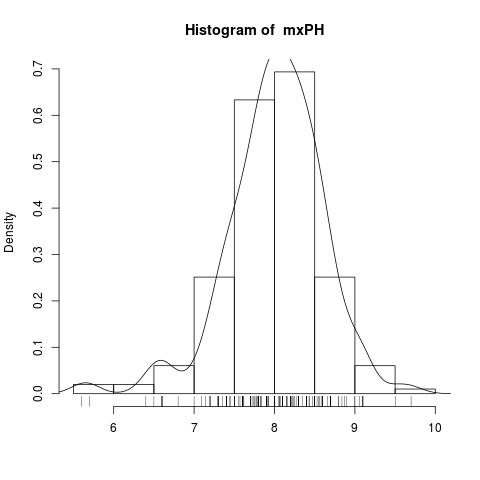
## Max. :31.600

##

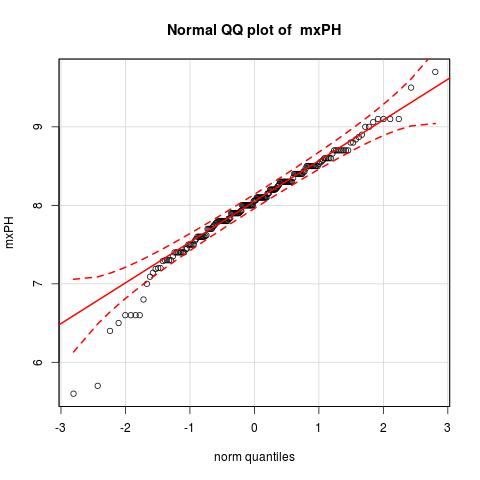
1. 数据可视化

2.1 数值数据

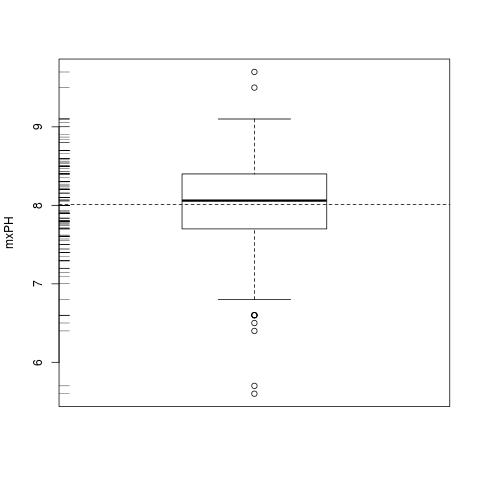
绘制直方图实例：mxPH



绘制Q-Q图实例：mxPH

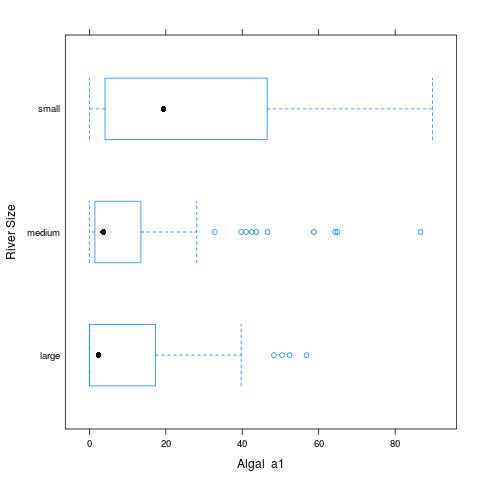


绘制盒图实例：mxPH，同时检测出离群点。

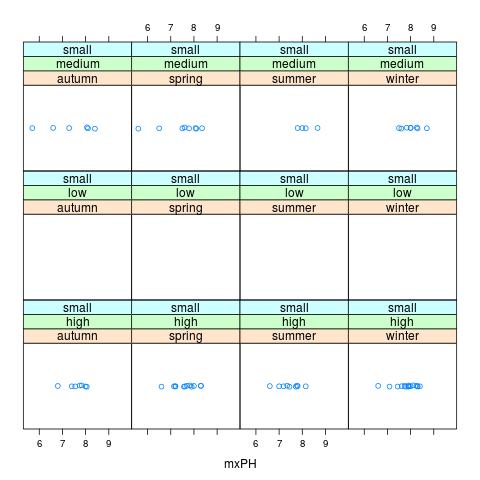


2.2 针对海藻

描述海藻1在不同size的河流中数量的盒图和离群点：



mxPH的条件箱图：

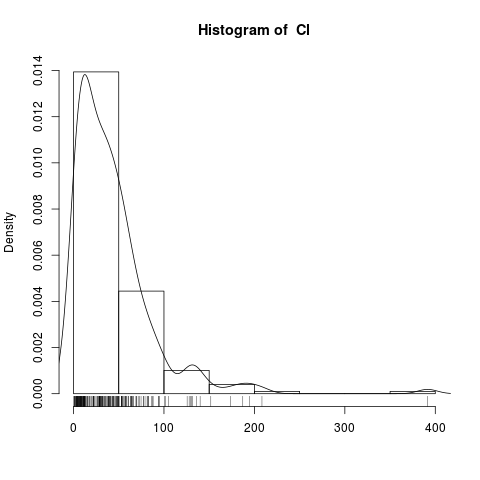
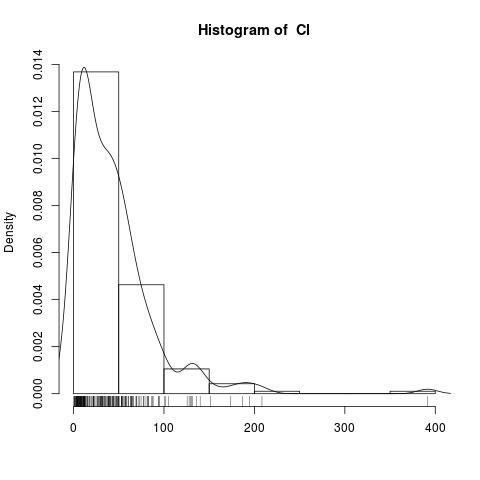


其他图像见仓库内的文件。

1. 缺失数据处理

我们对缺失数据进行了四种处理方法：1.直接删除该样本，2.用出现频率最大的值代替，3.通过数据集列（变量）相关关系，用相关度最大的其他数据代替，4.通过观测值之间的相关性。

对于缺失数据较多的Chla这个属性，下图举例了用方法3填充的之后（右）比填充之前（左）更加平滑。



其他图片参见仓库。