# 海藻数据的分析

## 环境搭建

1. 安装R语言软件包
2. 安装DMwR、car等进行该实验数据分析必要的程序包（如果报错，那就收到下载程序包，解压缩释放到R软件的Library文件夹下）

## 实验过程

**数据的可视化与摘要**

1. 初始化数据

代码如下，需要引入两个库：DMwR、car

把Analysis.txt的数据读入hz\_Analysis中

|  |
| --- |
| #初始化  library(DMwR)  library(car)  #读取数据文件  hz\_Analysis<-read.table(file,,col.names=c('season','riverSize','waterSpeed','maxPH','minO2','avrCL','avrNO3','avrNH4','avrPO43','avrPO4','avrYLS','a1','a2','a3','a4','a5','a6','a7'),na.strings=c('XXXXXXX')) |

运行结果如图：

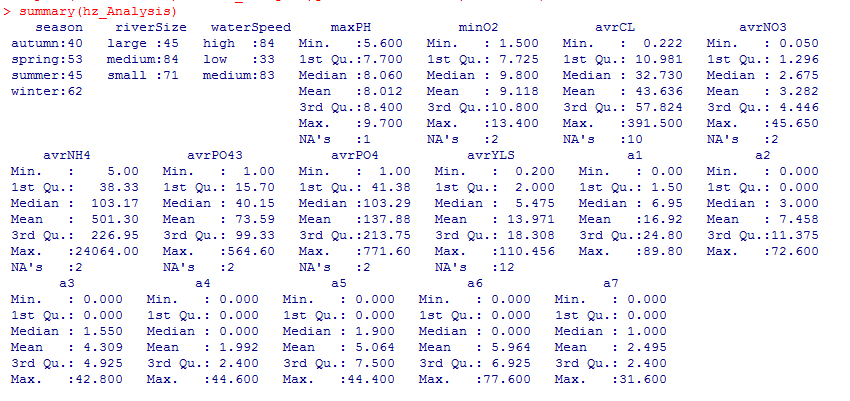


1. 数据摘要

直接在console里面输入命令：

|  |
| --- |
| summary(hz\_Analysis) |

结果如图：

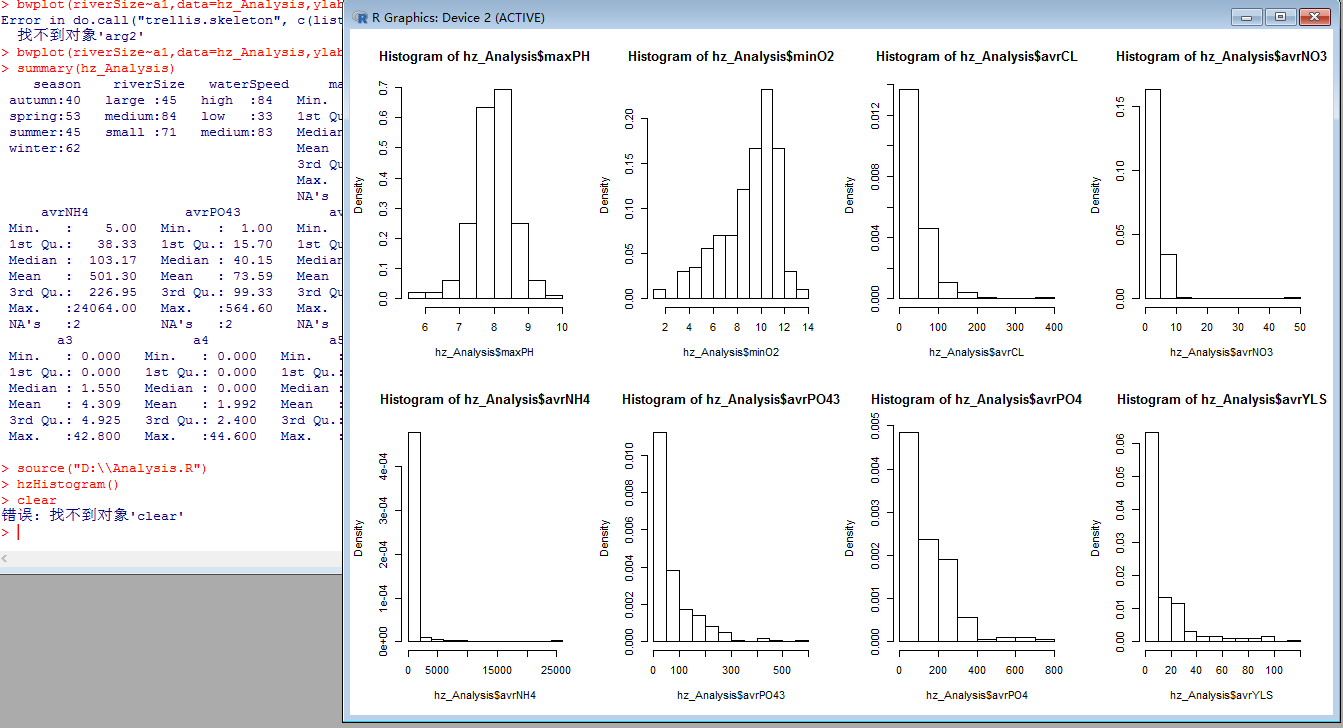


1. 画出直方图

代码如下，直接根据水中的各类统计数据画出直方图

|  |
| --- |
| #直方图  hzHistogram<-function()  {  par(mfrow=c(2,4))  hist(hz\_Analysis$maxPH,probability=T)  hist(hz\_Analysis$minO2,probability=T)  hist(hz\_Analysis$avrCL,probability=T)  hist(hz\_Analysis$avrNO3,probability=T)  hist(hz\_Analysis$avrNH4,probability=T)  hist(hz\_Analysis$avrPO43,probability=T)  hist(hz\_Analysis$avrPO4,probability=T)  hist(hz\_Analysis$avrYLS,probability=T)  } |

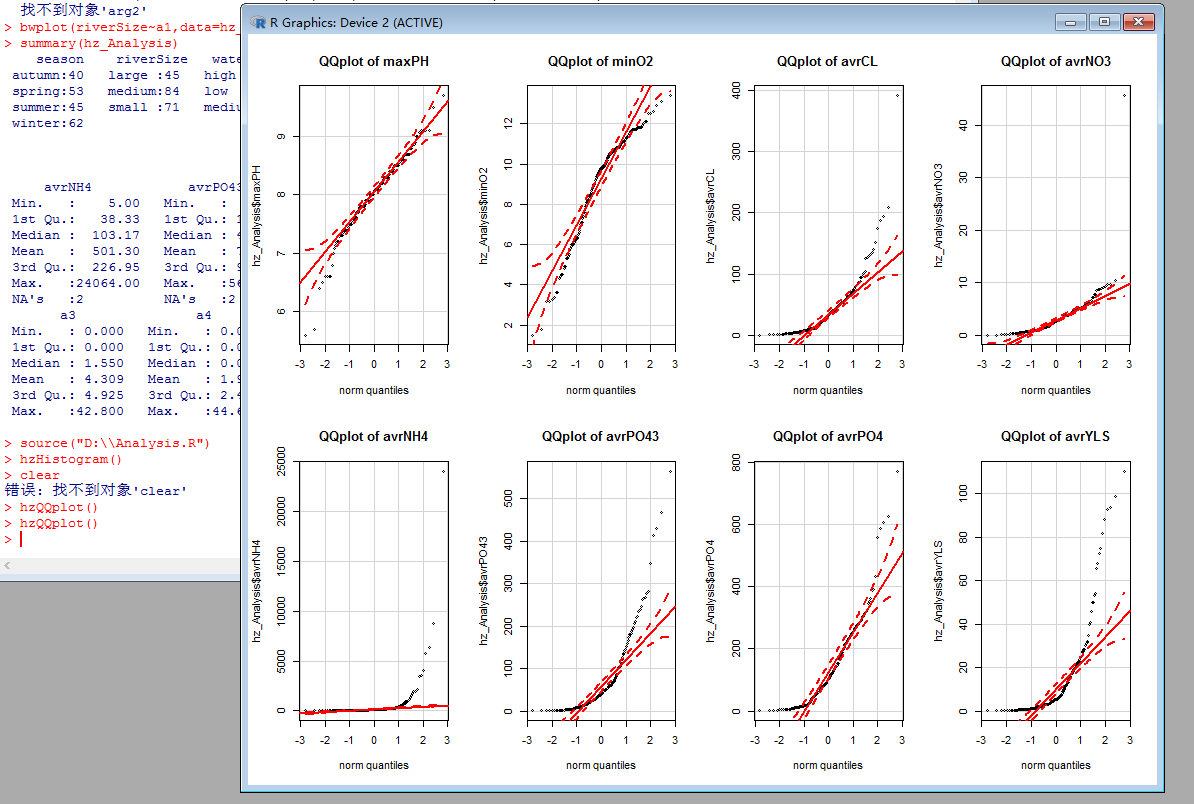
结果如图：



1. 画qq图

|  |
| --- |
| #qq图验证正态分布  hzQQplot<-function()  {  par(mfrow=c(2,4))  qqPlot(hz\_Analysis$maxPH, main='QQplot of maxPH')  qqPlot(hz\_Analysis$minO2, main='QQplot of minO2')  qqPlot(hz\_Analysis$avrCL, main='QQplot of avrCL')  qqPlot(hz\_Analysis$avrNO3, main='QQplot of avrNO3')  qqPlot(hz\_Analysis$avrNH4, main='QQplot of avrNH4')  qqPlot(hz\_Analysis$avrPO43, main='QQplot of avrPO43')  qqPlot(hz\_Analysis$avrPO4, main='QQplot of avrPO4')  qqPlot(hz\_Analysis$avrYLS, main='QQplot of avrYLS')  } |

结果如图：

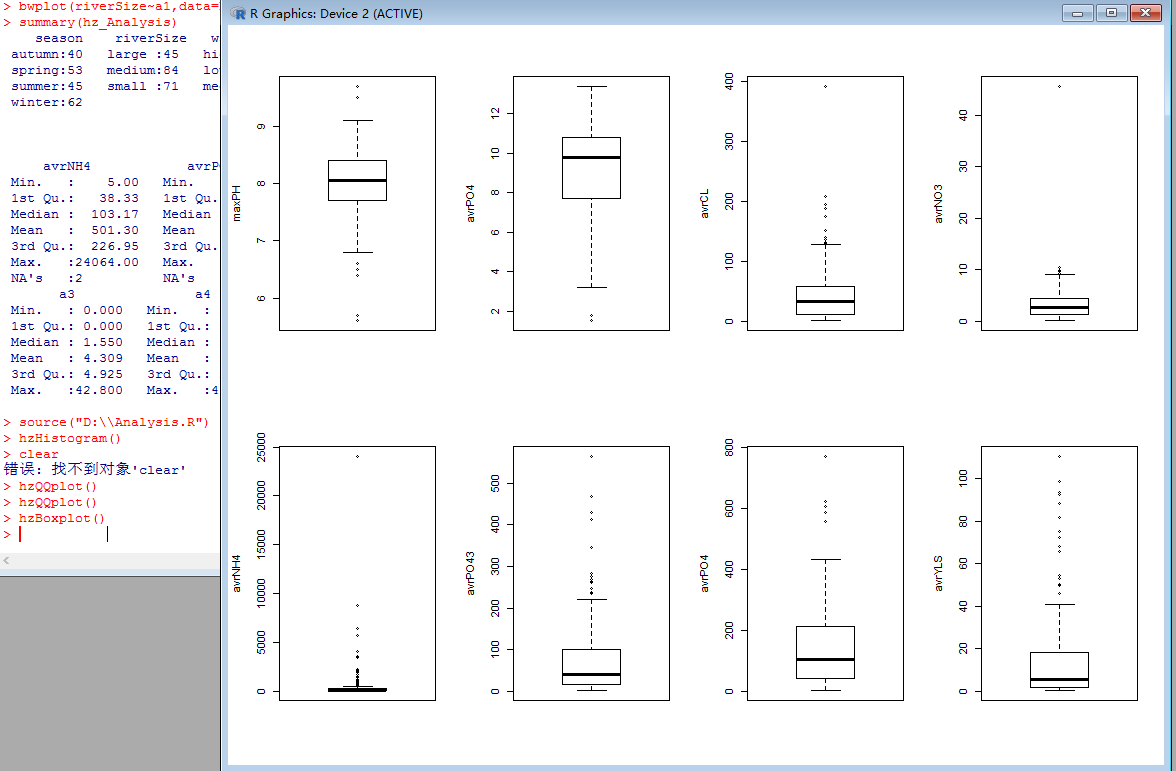


可以看出，只有maxPH的图比较接近正态分布

1. 画出盒图

|  |
| --- |
| #单独的盒图  hzBoxplot<-function()  {  par(mfrow=c(2,4))  boxplot(hz\_Analysis$maxPH,ylab="maxPH")  boxplot(hz\_Analysis$minO2,ylab="avrPO4")  boxplot(hz\_Analysis$avrCL,ylab="avrCL")  boxplot(hz\_Analysis$avrNO3,ylab="avrNO3")  boxplot(hz\_Analysis$avrNH4,ylab="avrNH4")  boxplot(hz\_Analysis$avrPO43,ylab="avrPO43")  boxplot(hz\_Analysis$avrPO4,ylab="avrPO4")  boxplot(hz\_Analysis$avrYLS,ylab="avrYLS")  } |

结果如图：

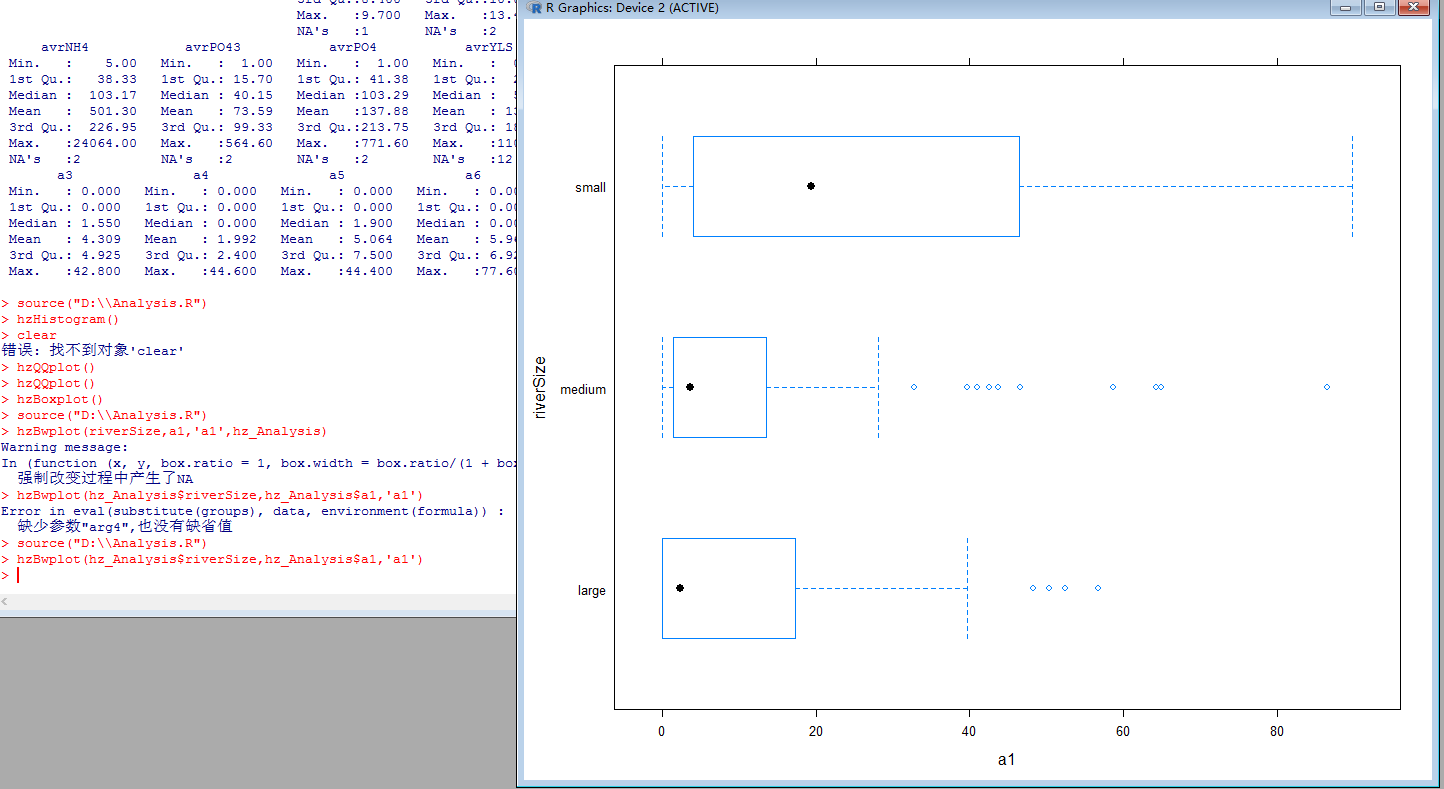


1. 画出条件盒图

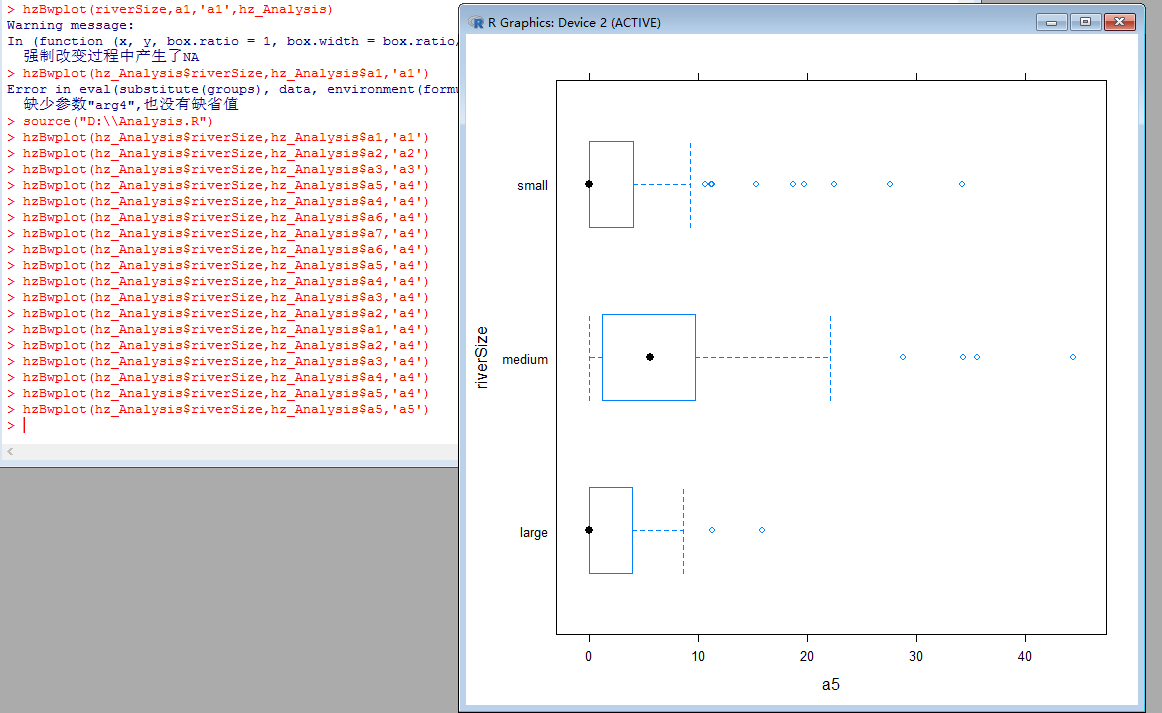
|  |
| --- |
| #条件盒图  hzBwplot<-function(arg1,arg2)  {  bwplot(arg1~arg2,data=hz\_Analysis,ylab = 'Y',xlab='X' )  } |

参数稍微说明下，arg1为riverSize或者是waterSpeed，arg2为a1-a7

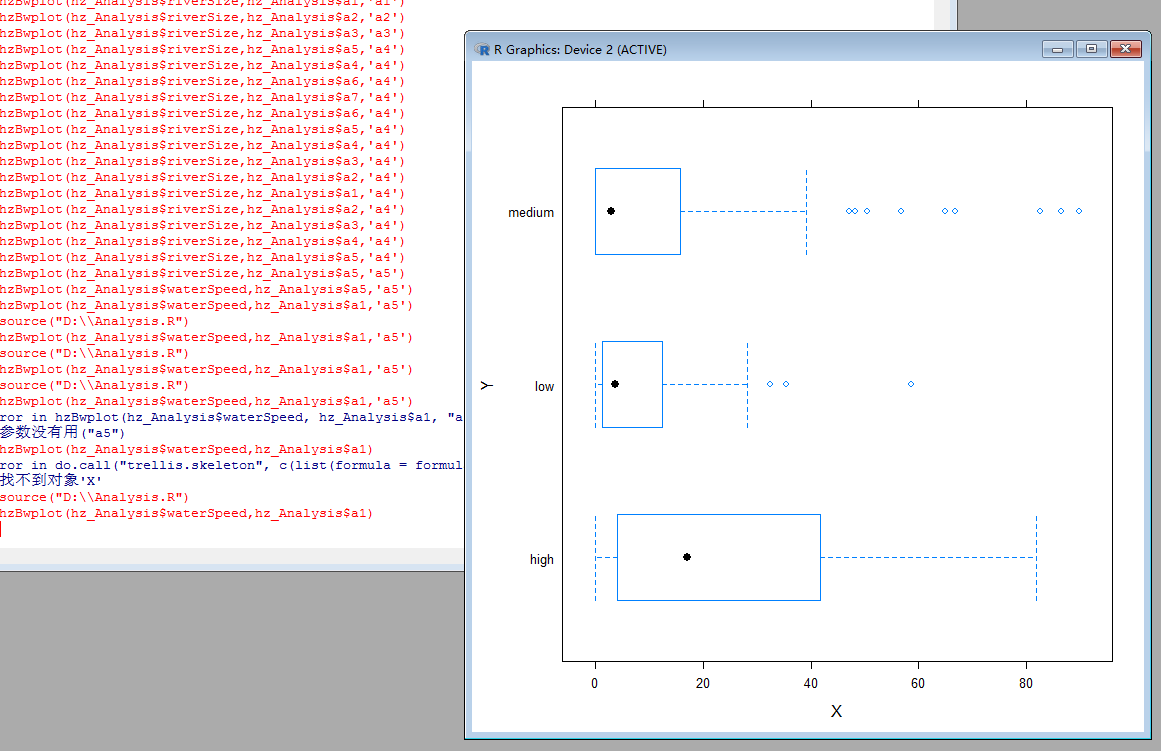
结果如图：



很容易看出，riverSize为small的河中，a1藻类频率较大



而a5则是riverSize为medium，频率较大

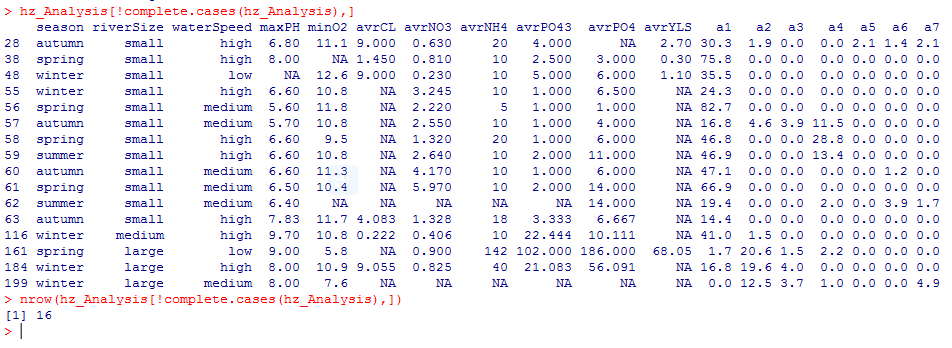


这个图则表明waterSpeed为high，a1的频率比较大。

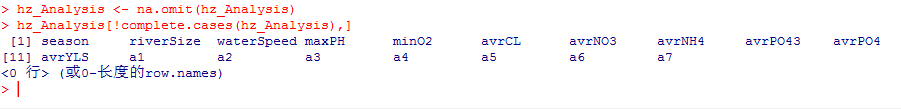
**数据缺失的处理**

1. 将缺失部分剔除

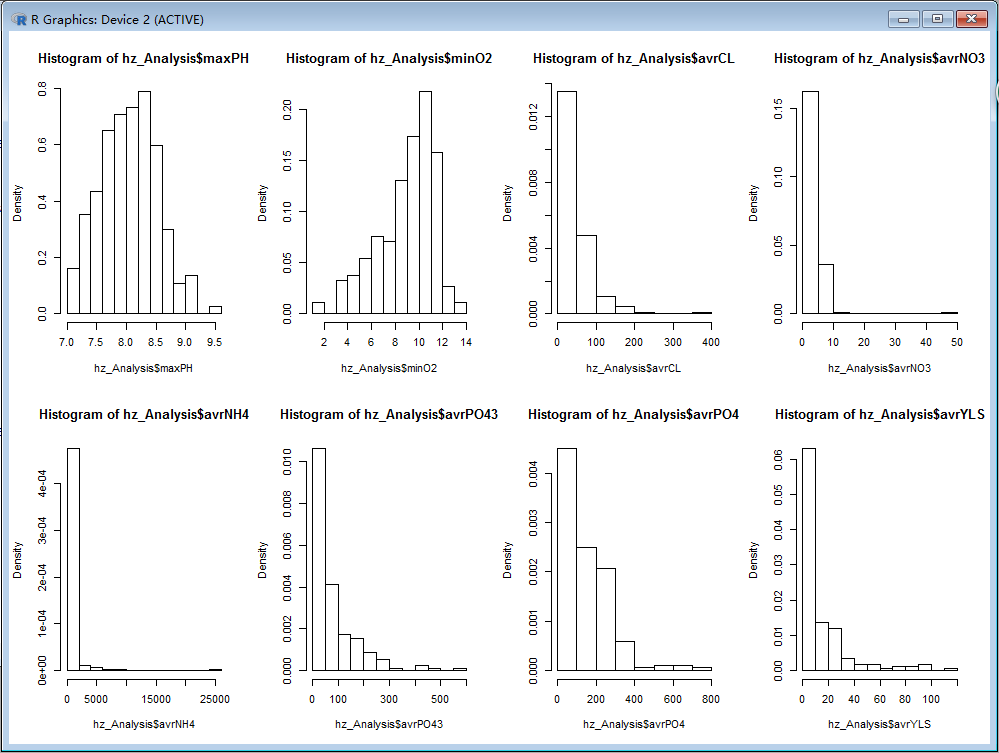
先查看缺失数据的条目：

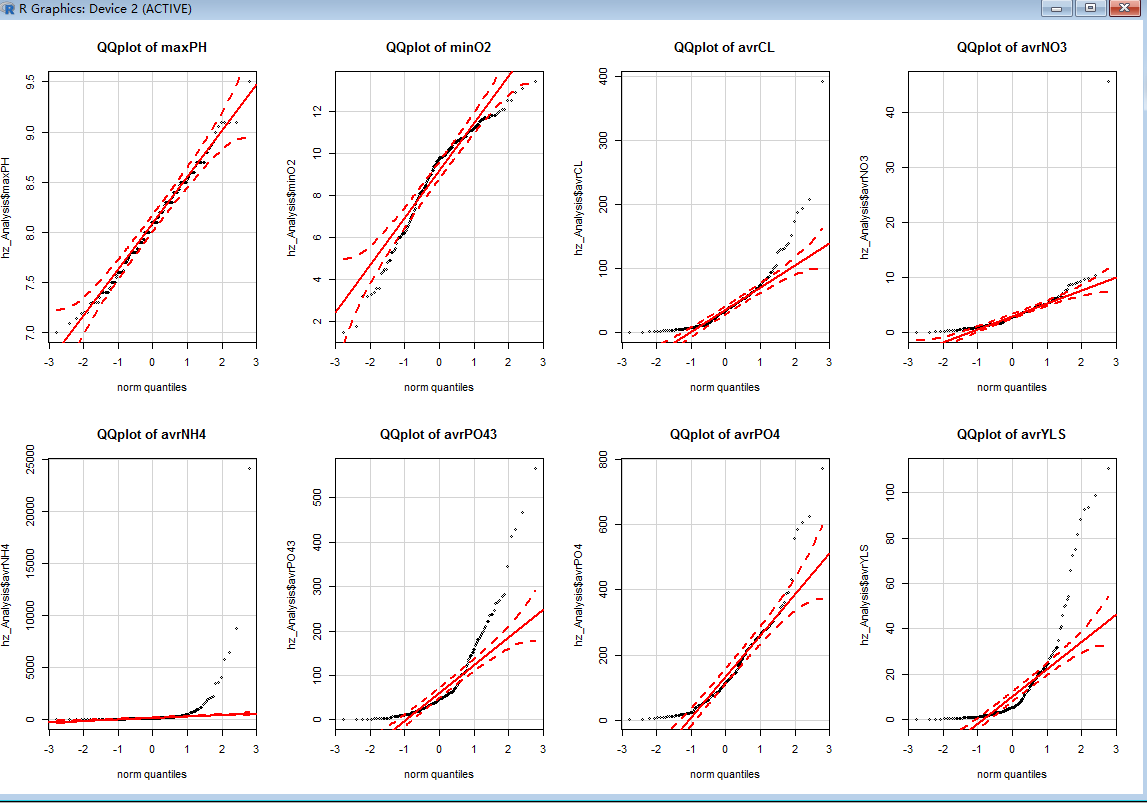


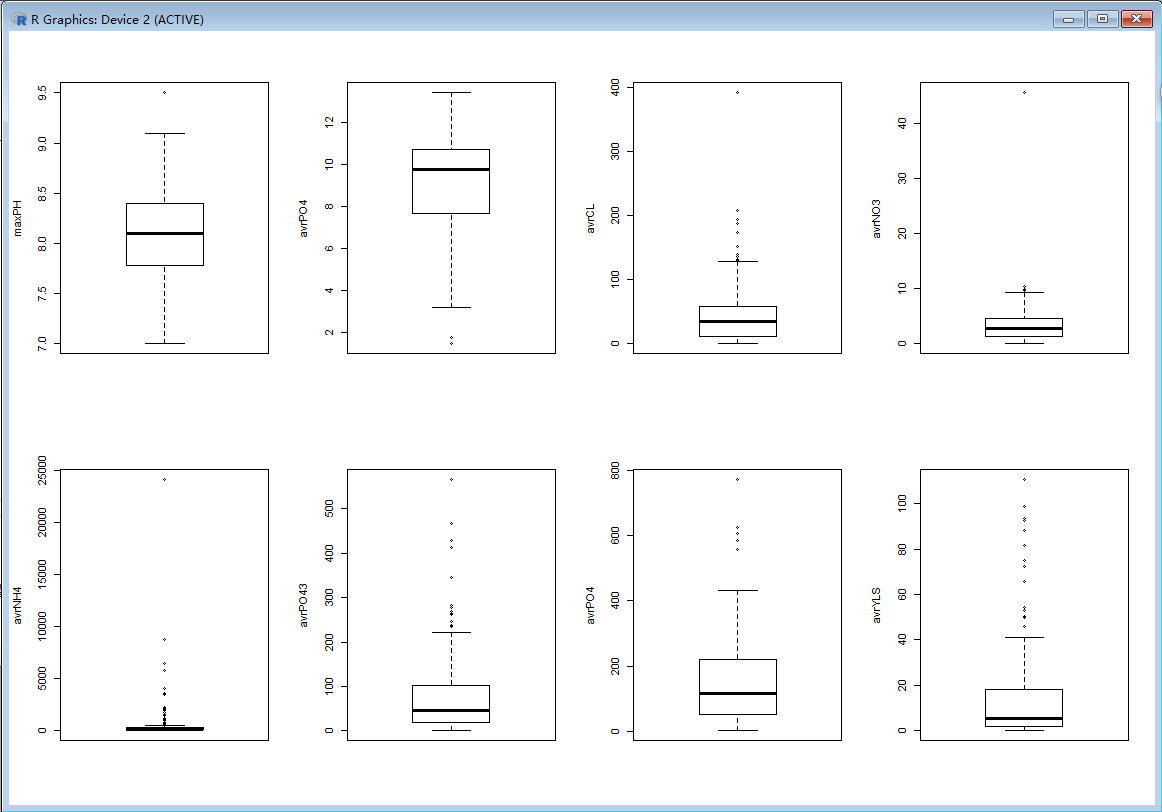
总共为16条，在200样本中，去掉影响也不太大，剔除掉



再看看直方图、qq图与盒图：



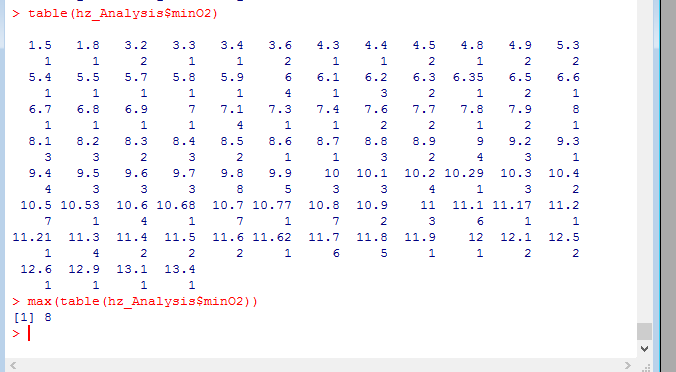




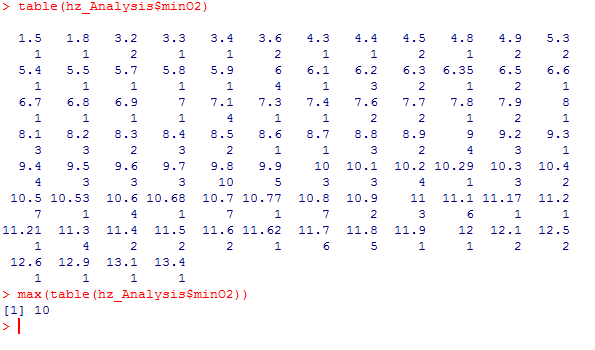
1. 用最高频率值来填补缺失值

重新加载数据，以minO2为例，

先求出最高频率的值：

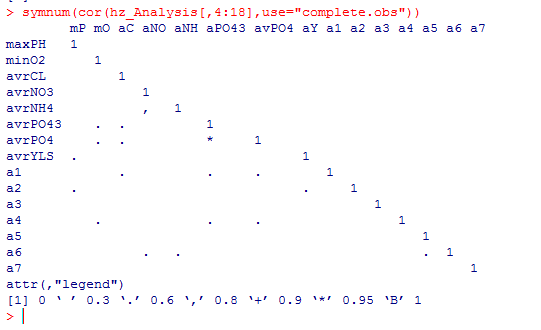


求出来最高频率值为9.8，替换掉之后为：



1. 通过属性的相关关系来填补缺失

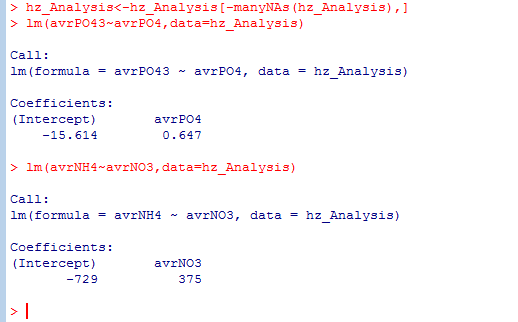
重新加载数据集，做相关系值矩阵分析



发现avrPO43与avrPO4的相关性相当高，他们两个可以相互补值

AvrNH4与avrNO3相关性也比较打，也可互相补值。

然后求取他们之间的函数关系：

avrPO43 = 0.647avrPO4 – 15.614

AvrNH4 = 375avrNO3 - 729

1. 通过数据对象之间的相似性来填补缺失值

重新加载数据



knnImputation(hz\_Analysis,k=10)，找出与NA数据中欧氏距离最近的的10个样本，用这10个样本的中位数来填充NA数据。

## 总结

第一、二种方法补全数据执行速度快，适用于大数据集，但是在一些小数据集上它可能导致较大的数据偏差，影响后期的数据分析工作。比如现在的200个数据集，就可能产生比较大的影响。

第三、四中方法补全数据误差比较小，但不全速度十分慢，适用于小数据集。

这次作业参考百度文库上几篇关于海藻数量数据处理的文章：

<http://wenku.baidu.com/link?url=HPmsjmPLV24qgYEy_mDFCdUI-6mIqfmvwaAmhgNV0sZ0BtN-qHfmIE-z-rOrBfDdX7nqpXEQLyfhOBZh9Vfg2vaTaYrD0GRzwHorvWcKB2_>

<http://wenku.baidu.com/view/5ef5e981f78a6529647d539d.html>

自己对R语言编程不是很熟，通过这次作业，对其有了初步的了解，同时也对数据分析、数据处理以及数据挖掘有了进一步直观的认识。