# 海藻数据的分析

**计算机学院 2120150980 谌重喜**

**一、实验要求**

### **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**

### 数据: Analysis.R

### 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。

## 二、实验环境

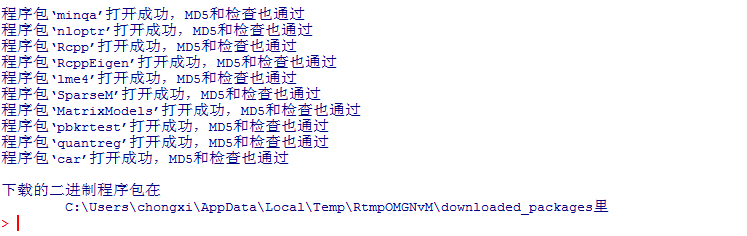
1. Win7操作系统、R语言软件包、DMwR、car等程序包

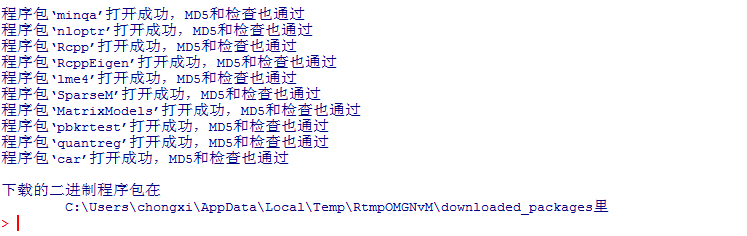
## 三、实验过程

**数据的可视化与摘要**

1. 实验准备

导入两个程序包：DMwR、car





2、读入数据

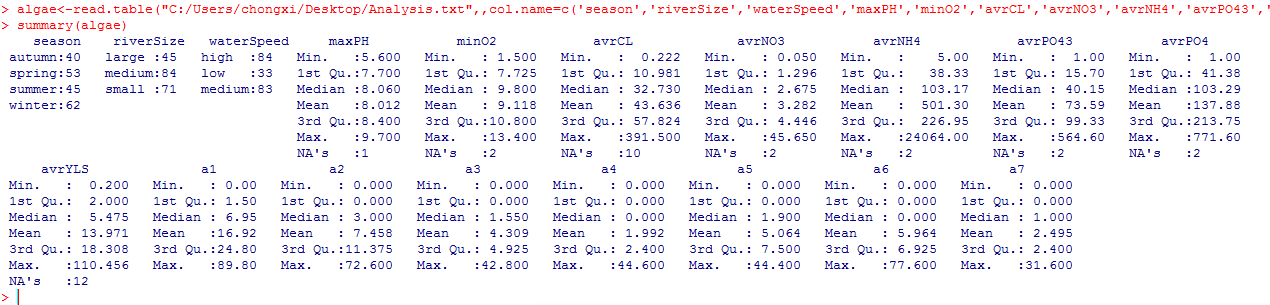
代码：

algae<-read.table("C:/Users/chongxi/Desktop/Analysis.txt",,col.name=c('season','riverSize',' waterSpeed','maxPH','minO2','avrCL','avrNO3','avrNH4','avrPO43','avrPO4','avrYLS','a1','a2','a3','a4','a5','a6','a7'),na.strings=c('XXXXXXX'))

summary(algae)

数据摘要

结果如图：

 3直方图

algaeHistogram<-function()

{

par(mfrow=c(2,4))

hist(algae$maxPH,probability=T)

hist(algae$minO2,probability=T)

hist(algae$avrCL,probability=T)

hist(algae$avrNO3,probability=T)

hist(algaes$avrNH4,probability=T)

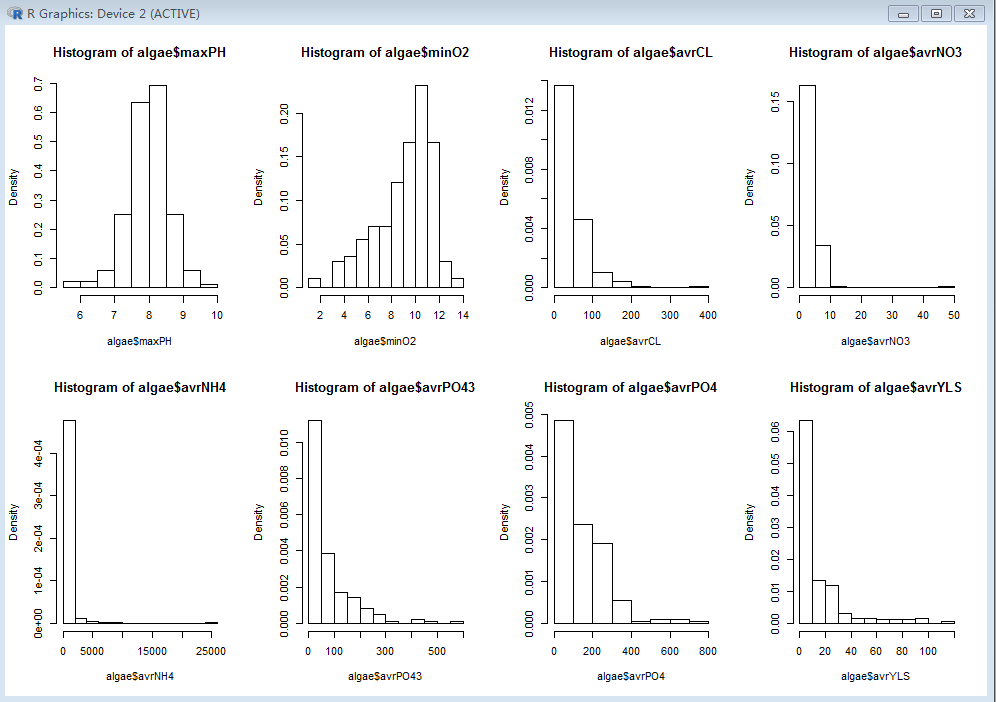
hist(algae$avrPO43,probability=T)

hist(algae$avrPO4,probability=T)

hist(algae$avrYLS,probability=T)

}

结果如图：



1. qq图

algaeQQplot<-function()

{

par(mfrow=c(2,4))

qqPlot(algae$maxPH, main='QQplot of maxPH')

qqPlot(algae$minO2, main='QQplot of minO2')

qqPlot(algae$avrCL, main='QQplot of avrCL')

qqPlot(algae$avrNO3, main='QQplot of avrNO3')

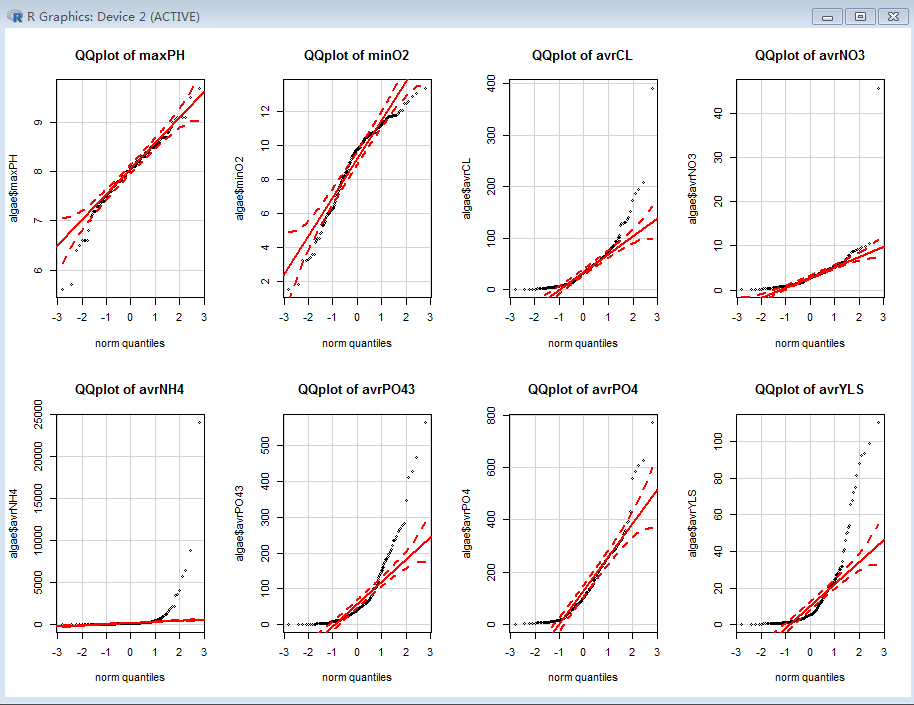
qqPlot(algae$avrNH4, main='QQplot of avrNH4')

qqPlot(algae$avrPO43, main='QQplot of avrPO43')

qqPlot(algae$avrPO4, main='QQplot of avrPO4')

qqPlot(algae$avrYLS, main='QQplot of avrYLS')

} 结果如图：



只有maxPH的图接近正态分布

5、盒图

algaeBoxplot<-function()

{

par(mfrow=c(2,4))

boxplot(algae$maxPH,ylab="maxPH")

boxplot(algae$minO2,ylab="avrPO4")

boxplot(algae$avrCL,ylab="avrCL")

boxplot(algae$avrNO3,ylab="avrNO3")

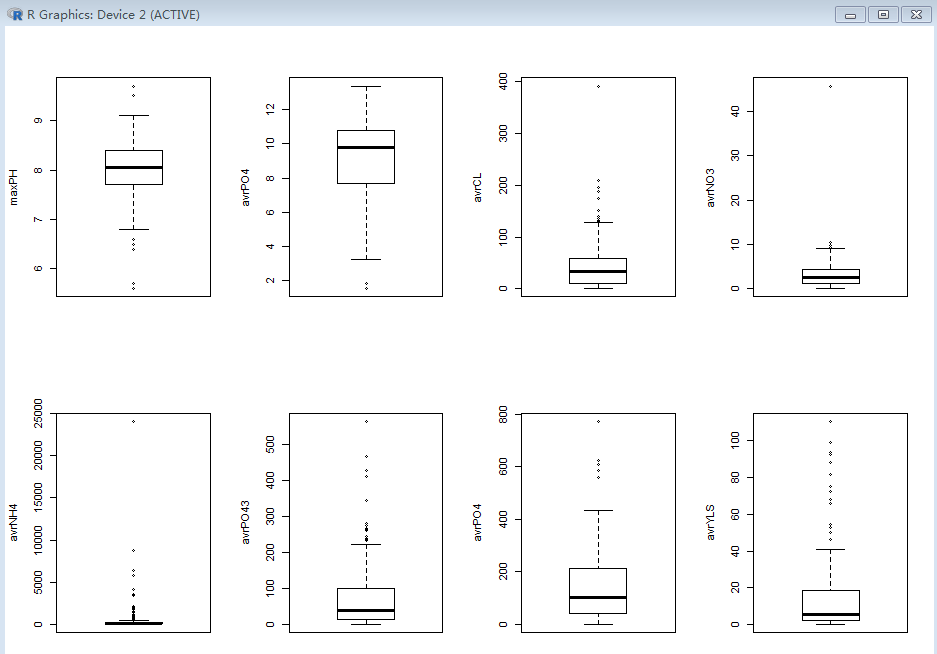
boxplot(algae$avrNH4,ylab="avrNH4")

boxplot(algae$avrPO43,ylab="avrPO43")

boxplot(algae$avrPO4,ylab="avrPO4")

boxplot(algae$avrYLS,ylab="avrYLS")

} 结果如图：



6、条件盒图

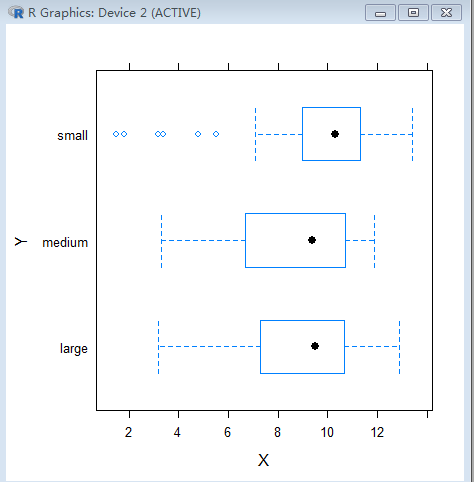
algaeBwplot<-function(arg1,arg2)

{

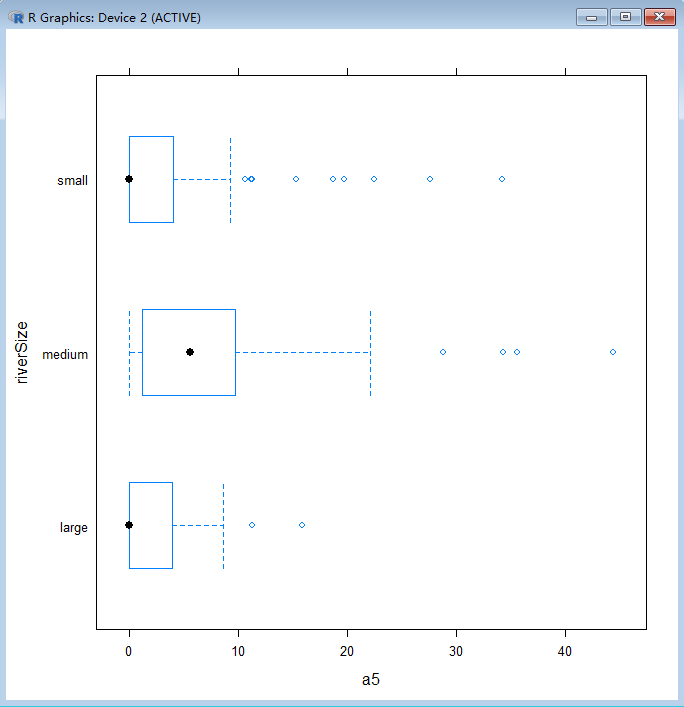
bwplot(arg1~arg2,data=algae,ylab = 'Y',xlab='X' )

}

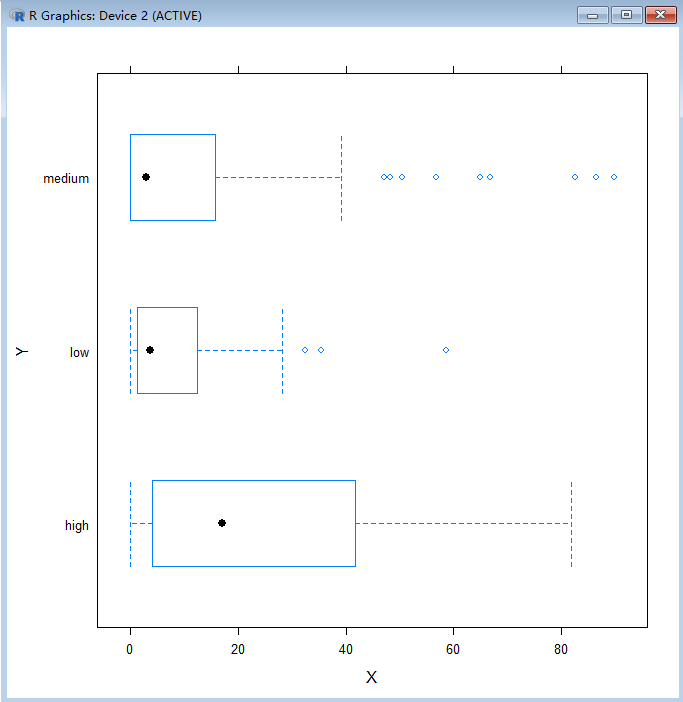
结果如图：



riverSize为small的河中，a1藻类频率较大



而a5则是riverSize为medium，频率较大

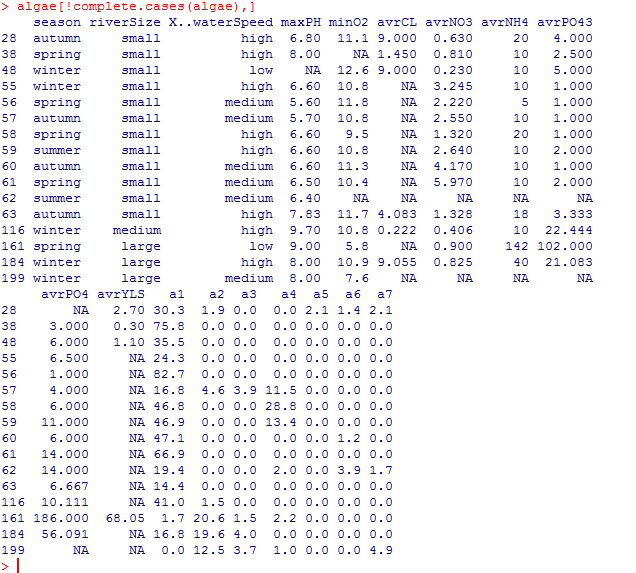


这个图则表明waterSpeed为high，a1的频率比较大。

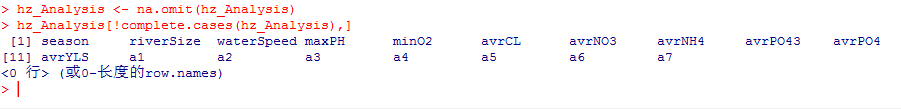
**数据缺失的处理**

1. 将缺失部分剔除

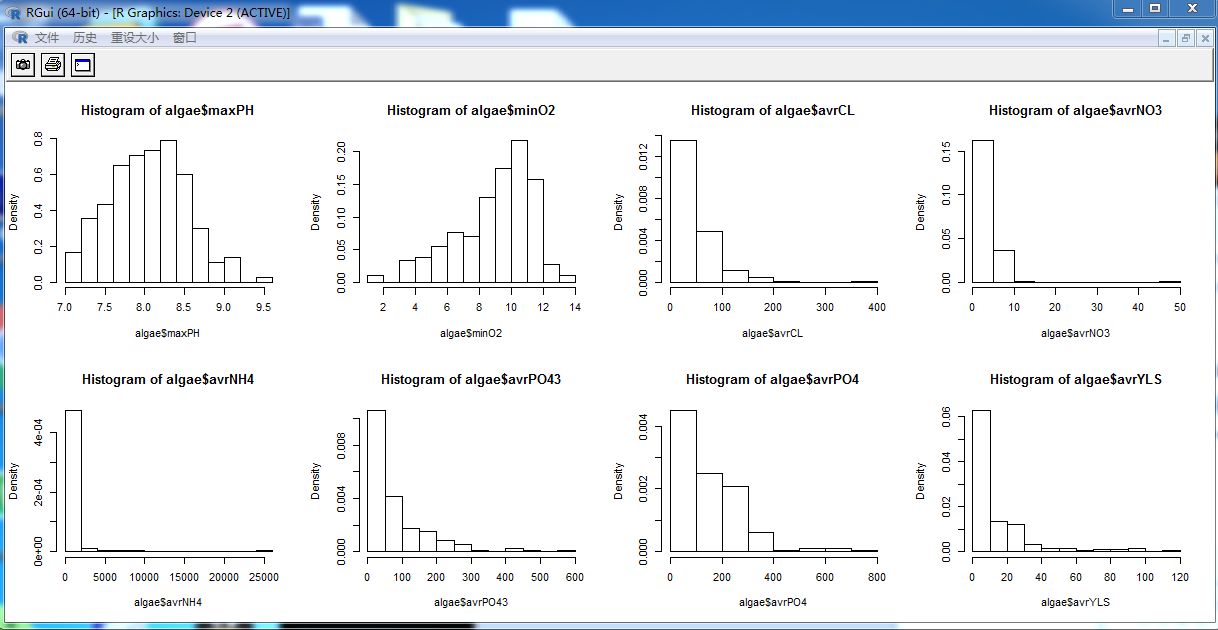
查看缺失数据的条目：

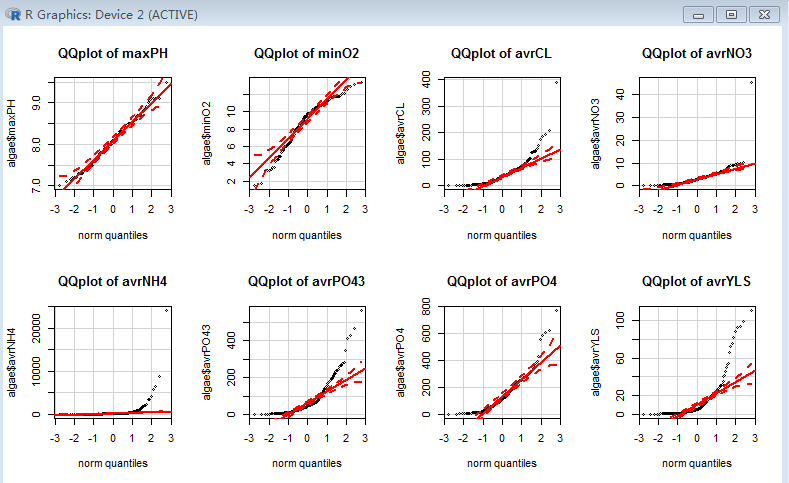


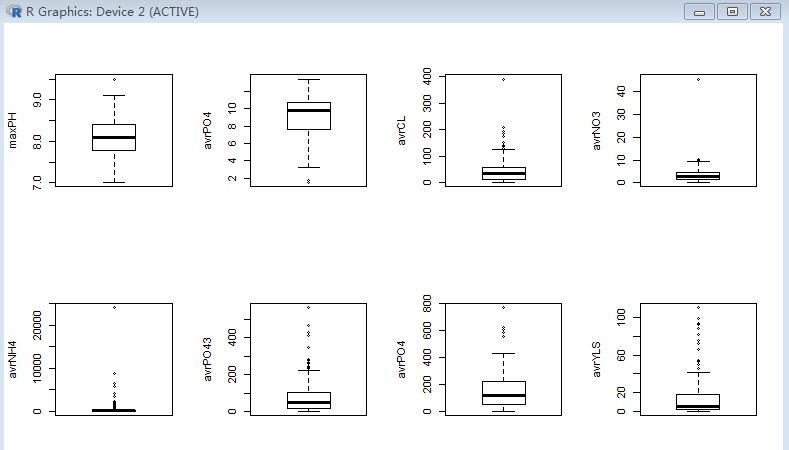
在200样本中，总共为16条，剔除



直方图、qq图与盒图：



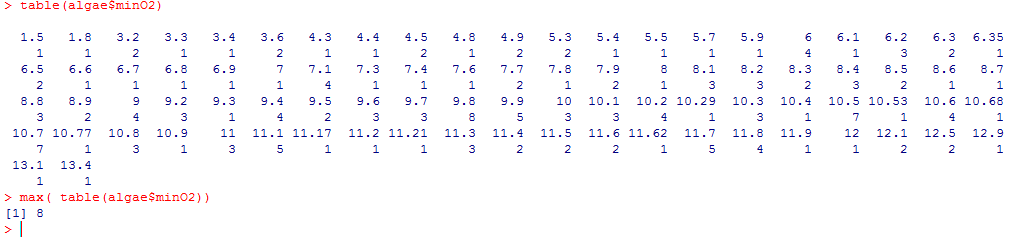




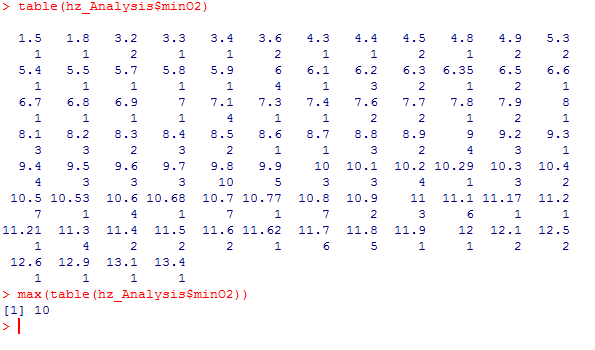
1. 通过最高频率值来填补缺失值

重新加载数据，

以minO2为例，求出最高频率值：

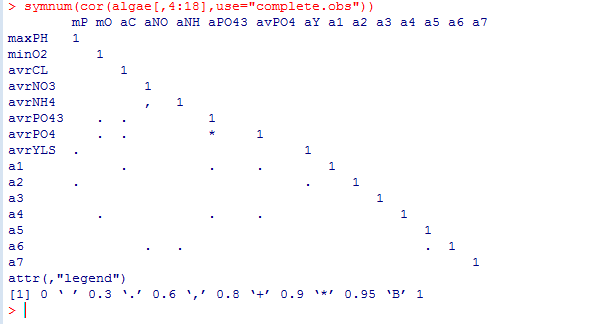


最高频率值为9.8，替换后为：



1. 通过属性的相关关系来填补缺失

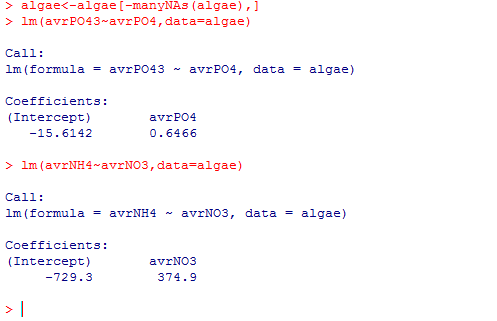
重新加载数据集，做相关系值矩阵分析



avrPO43与avrPO4相关性高，可以相互补值

AvrNH4与avrNO3相关性比较高，也可互相补值。

求取其之间的函数关系：

avrPO43 = 0.6466 avrPO4 = – 15.6142

AvrNH4 = 374.9 avrNO3 = – 729.3

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

53C5.tmp

knnImputation(algae,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。

## 四、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、四种方法补全数据误差比较小，但不全速度慢，适用于小数据集。

通过这次作业，对R语言编程有一定了解，对数据分析、数据处理以及数据挖掘有初步认识。

1. **代码附录**

### #初始化

### library(DMwR)

### library(car)

### #读取数据文件

### algae<-read.table("C:/Users/chongxi/Desktop/Analysis.txt",,col.name=c('season','riverSize',' waterSpeed','maxPH','minO2','avrCL','avrNO3','avrNH4','avrPO43','avrPO4','avrYLS','a1','a2','a3','a4','a5','a6','a7'),na.strings=c('XXXXXXX'))

### #摘要

### #summary(algae)

### #直方图

### algaeHistogram<-function()

### {

### par(mfrow=c(2,4))

### hist(algae$maxPH,probability=T)

### hist(algae$minO2,probability=T)

### hist(algae$avrCL,probability=T)

### hist(algae$avrNO3,probability=T)

### hist(algae$avrNH4,probability=T)

### hist(algae$avrPO43,probability=T)

### hist(algae$avrPO4,probability=T)

### hist(algae$avrYLS,probability=T)

### }

### #qq图验证正态分布

### algaeQQplot<-function()

### {

### par(mfrow=c(2,4))

### qqPlot(algae$maxPH, main='QQplot of maxPH')

### qqPlot(algae$minO2, main='QQplot of minO2')

### qqPlot(algae$avrCL, main='QQplot of avrCL')

### qqPlot(algae$avrNO3, main='QQplot of avrNO3')

### qqPlot(algae$avrNH4, main='QQplot of avrNH4')

### qqPlot(algae$avrPO43, main='QQplot of avrPO43')

### qqPlot(algae$avrPO4, main='QQplot of avrPO4')

### qqPlot(algae$avrYLS, main='QQplot of avrYLS')

### }

### #单独的盒图

### algaeBoxplot<-function()

### {

### par(mfrow=c(2,4))

### boxplot(algae$maxPH,ylab="maxPH")

### boxplot(algae$minO2,ylab="avrPO4")

### boxplot(algae$avrCL,ylab="avrCL")

### boxplot(algae$avrNO3,ylab="avrNO3")

### boxplot(algae$avrNH4,ylab="avrNH4")

### boxplot(algae$avrPO43,ylab="avrPO43")

### boxplot(algae$avrPO4,ylab="avrPO4")

### boxplot(algae$avrYLS,ylab="avrYLS")

### }

### #条件盒图

### algaeBwplot<-function(arg1,arg2)

### {

### bwplot(arg1~arg2,data=algae,ylab = 'Y',xlab='X' )

### }

### #将缺失部分剔除

### #algae[!complete.cases(algae),]

### #nrow(algae[!complete.cases(algae),])

### #algae <- na.omit(algae)

### #用最高频率值来填补缺失值

### #table(a)

### #max(table(a))

### #algae[is.na(algae$minO2),'minO2'] <- 9.8

### #通过属性的相关关系来填补缺失

### #options(digits=1)

### #cor(algae[,4:18],use="complete.obs")

### #symnum(cor(algae[,4:18],use="complete.obs"))

### #data(algae)

### #algae<-algae[-manyNAs(algae),]

### #lm(avrPO43~avrPO4,data=algae)

### #通过数据对象之间的相似性来填补缺失值

### #algae<-algae[-manyNAs(algae),]

### #algae<-knnImputation(algae,k=10)

### #algae<-knnImputation(algae,k=10,meth="median")