**海藻数据的分析**

**1 数据摘要**

读取文件，第一个参数指向了待分析数据文件的路径位置, 同时设定表每列属性名，设定缺失数据的字符串为XXXXXXX。

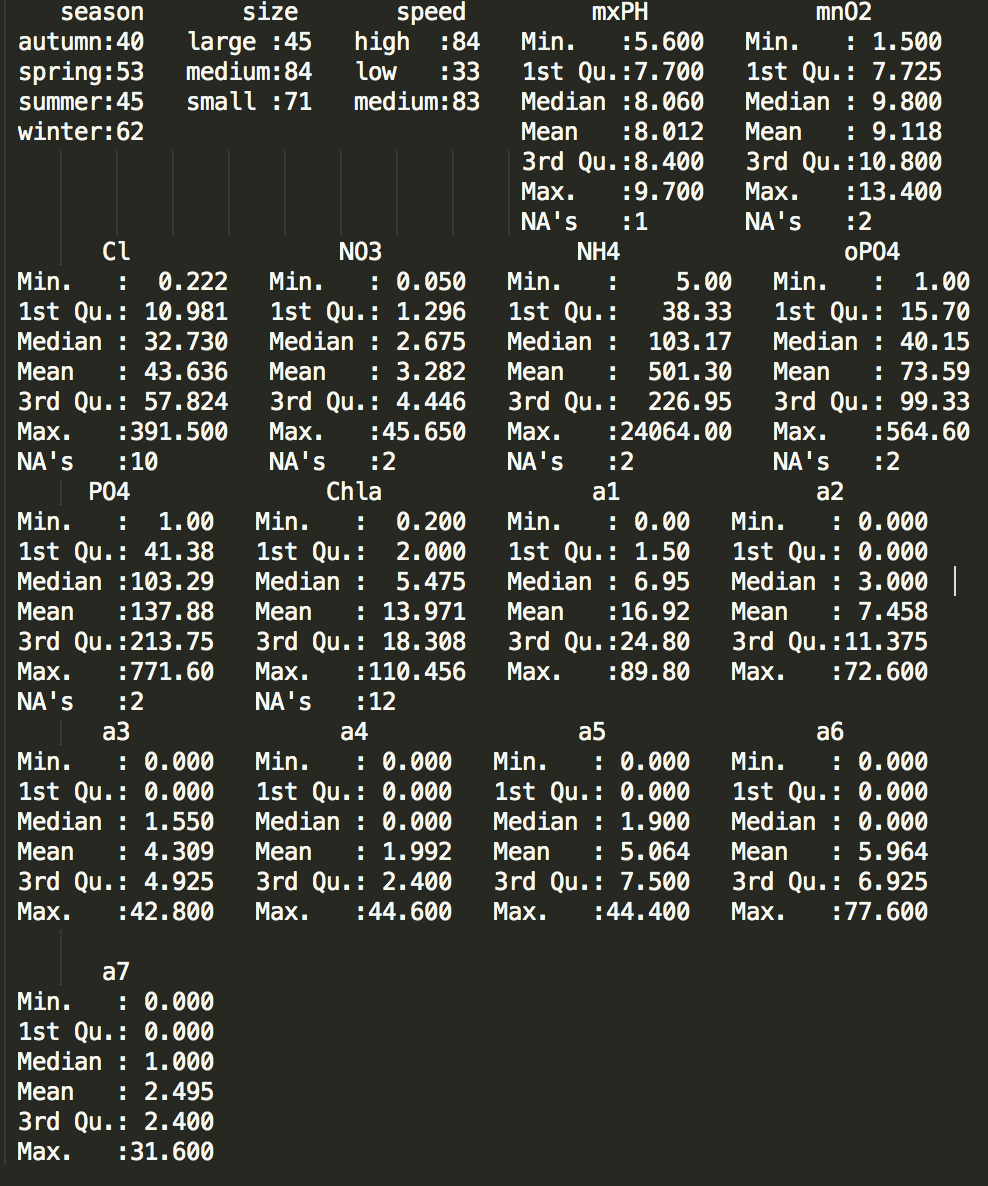
使用R代码：

location = "Analysis.txt"

data <- read.table(location,col.name = c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'),na.string=('XXXXXXX'))

用分析数据摘要使用summary：summary(data)

得到如下信息，最小值、前四分位，中位数，平均值，后四分位，最大值以及缺失数量的统计信息，如图：



**2 数据预处理以及对比可视化**

1. 对数值属性，绘制直方图与QQ图检验其正态分布

绘制其直方图与QQ图。绘制出的直方图纵轴是其频数，横轴是其分布区间。QQ图中，红色实线为其QQ线，虚线为95%置信度的置信区间。

draw <- function(list,list1,file,title)

{

ppath = file

png(file=ppath,width=1200,height=1000)

split.screen(c(3,2))

par(pin=c(4,4))

screen(1)

hist(list,main = paste(title," ",sep =""))

screen(3)

qqPlot(list,main=paste(title," ",sep =""))

screen(5)

boxplot(list, main=paste(title," ",sep =""), na.rm = TRUE)

screen(2)

hist(list1,main = paste(title," original",sep =""))

screen(4)

qqPlot(list1,main=paste(title," original",sep =""))

screen(6)

boxplot(list1, main=paste(title," original",sep =""), na.rm = TRUE)

dev.off()

}

函数中，list为数据预处理之后的某一列数据，list1为原始的同一列数据，file为保存图片的路径，title为图标题。

在每次执行数据预处理之后，调用下面的函数画出所有的对比图

draw\_all <- function(Adata,data,dir)

{

draw(Adata$mxPH,data$mxPH,paste(dir,"mxPH.png",sep =""),"mxPH")

draw(Adata$mnO2,data$mnO2,paste(dir,"mnO2.png",sep =""),"mnO2")

draw(Adata$Cl,data$Cl,paste(dir,"Cl.png",sep =""),"Cl")

draw(Adata$NO3,data$NO3,paste(dir,"NO3.png",sep =""),"NO3")

draw(Adata$NH4,data$NH4,paste(dir,"NH4.png",sep =""),"NH4")

draw(Adata$oPO4,data$oPO4,paste(dir,"oPO4.png",sep =""),"oPO4")

draw(Adata$PO4,data$PO4,paste(dir,"PO4.png",sep =""),"PO4")

draw(Adata$Chla,data$Chla,paste(dir,"Chla.png",sep =""),"Chla")

draw(Adata$a1,data$a1,paste(dir,"a1.png",sep =""),"a1")

draw(Adata$a2,data$a2,paste(dir,"a2.png",sep =""),"a2")

draw(Adata$a3,data$a3,paste(dir,"a3.png",sep =""),"a3")

draw(Adata$a4,data$a4,paste(dir,"a4.png",sep =""),"a4")

draw(Adata$a5,data$a5,paste(dir,"a5.png",sep =""),"a5")

draw(Adata$a6,data$a6,paste(dir,"a6.png",sep =""),"a6")

draw(Adata$a7,data$a7,paste(dir,"a7.png",sep =""),"a7")

}

数据预处理方法

（1） 剔除缺失值

clean\_NA <-function(Adata)

{

Adata[!complete.cases(Adata),]

Adata <- na.omit(Adata)

Adata

}

对比图如下：以mxPH为例，其他所有对比图都有生成，在clean\_NA文件夹下。

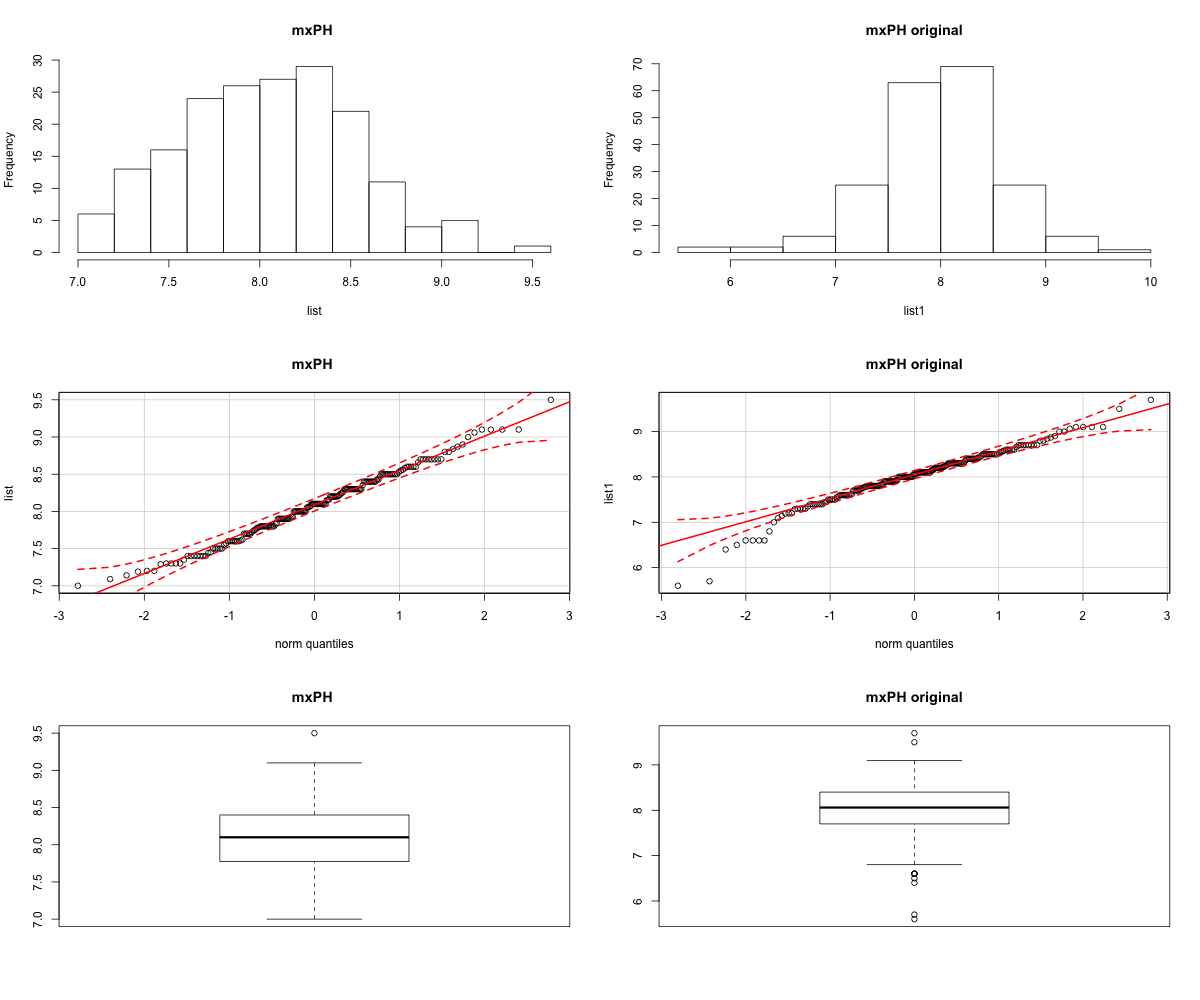


图1剔除缺失值之后对比图

左列为处理之后的，右列为原始数据的图，可以发现数据有所改变。

（2） 用最高频率值来填补缺失值

central\_NA <- function(Adata)

{

#data(Adata)

Adata <- Adata[-manyNAs(Adata)]

Adata <- centralImputation(Adata)

Adata

}

对比图如下：以mxPH为例，其他所有对比图都有生成，在central\_NA文件夹下。

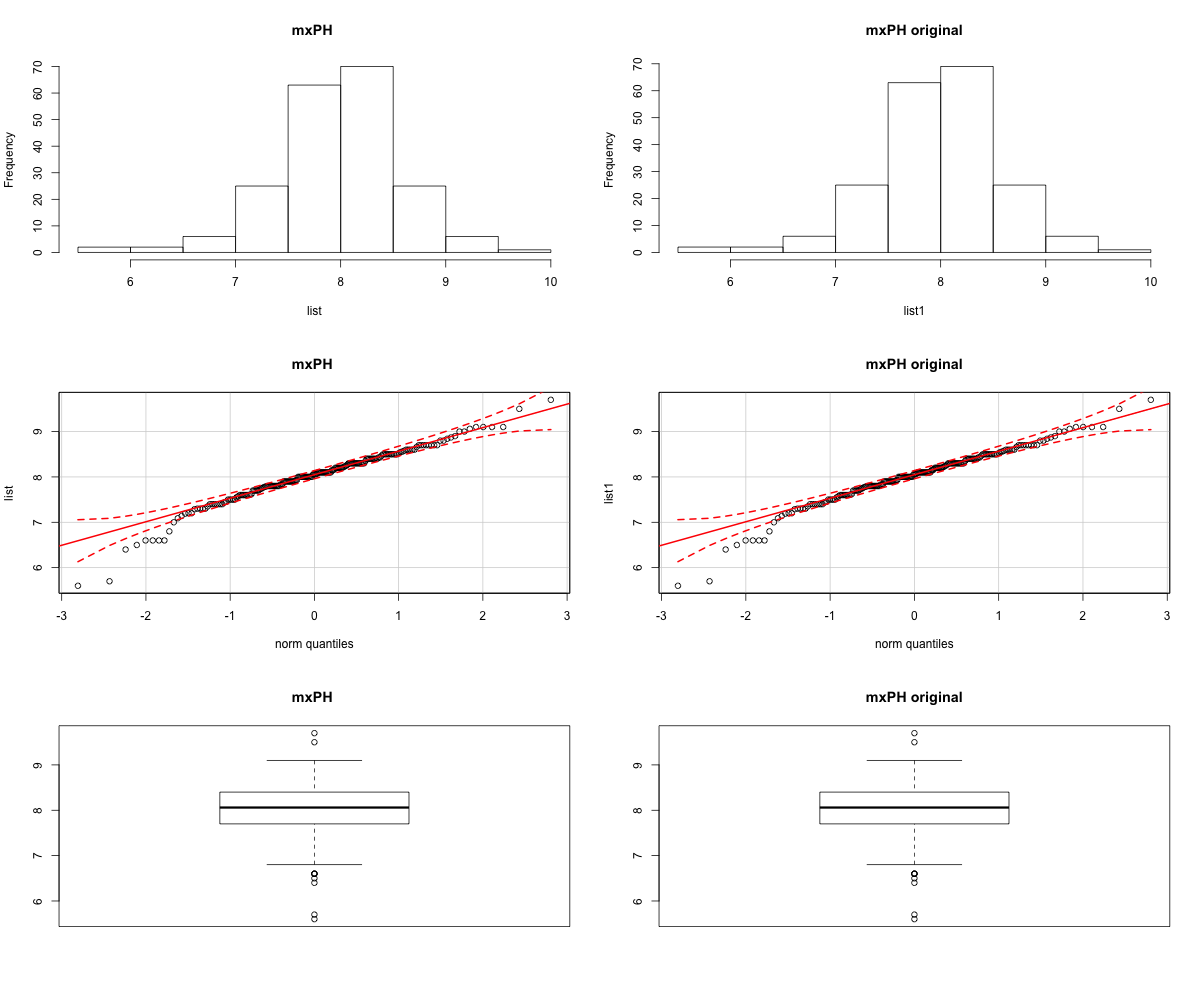


图2用最高频率值来填补缺失值对比图

可以看出，数据被很好的还原，保存了其主要特性。

* （3）通过属性的相关关系来填补缺失值

relation\_NA <- function(Adata)

{

symnum(cor(Adata[,4:18],use='complete.obs'))

lm(formula=PO4~oPO4, data=Adata)

Adata = Adata[-manyNAs(Adata),]

Adata

}

对比图如下：以mxPH为例，其他所有对比图都有生成，在relation\_NA文件夹

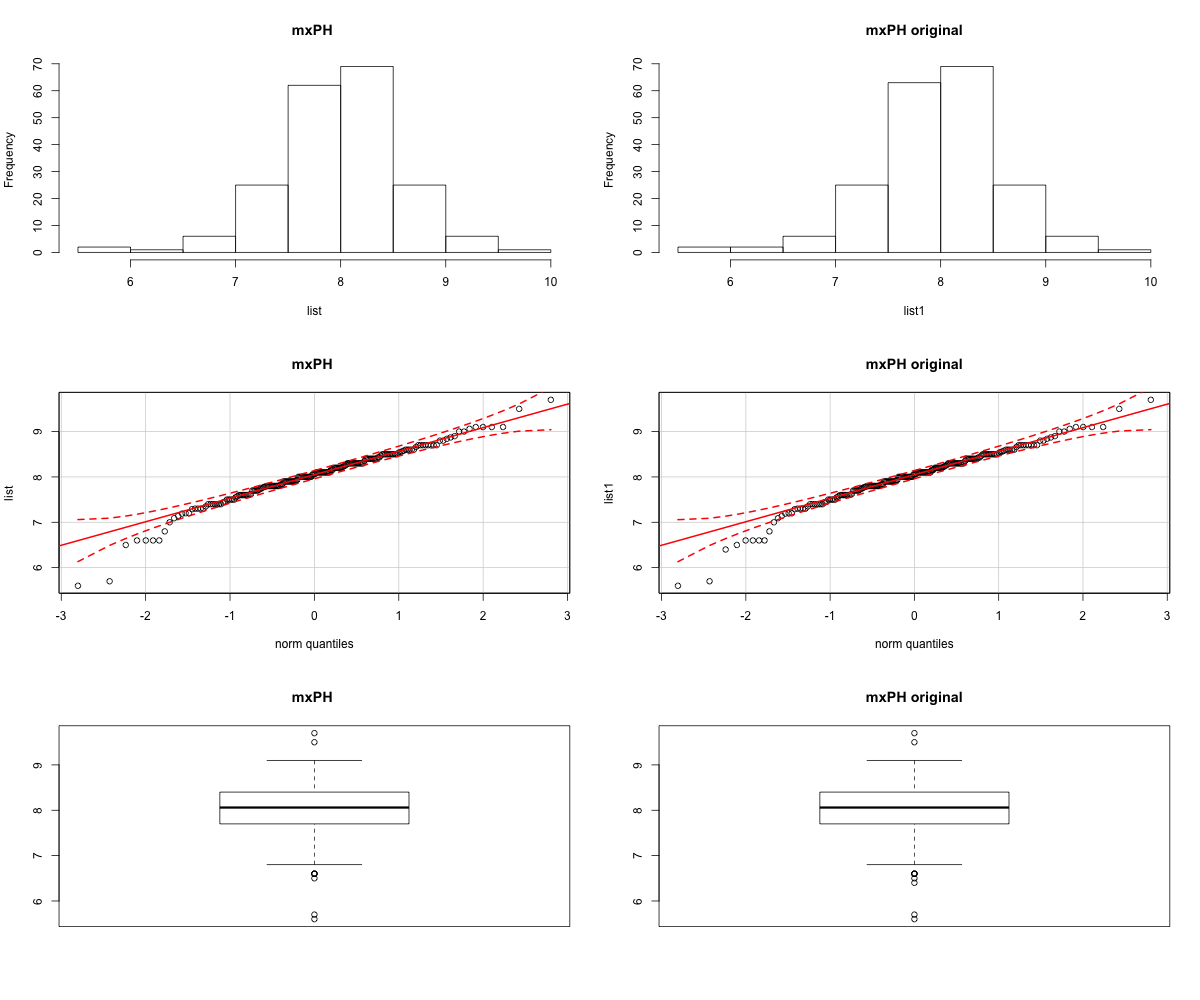


图3通过属性的相关关系来填补缺失值对比图

可以看出，数据被很好的还原，保存了其主要特性。

（4）通过数据对象之间的相似性来填补缺失值

knn\_NA <- function(Adata)

{

Adata <- knnImputation(Adata,10)

}

对比图如下：

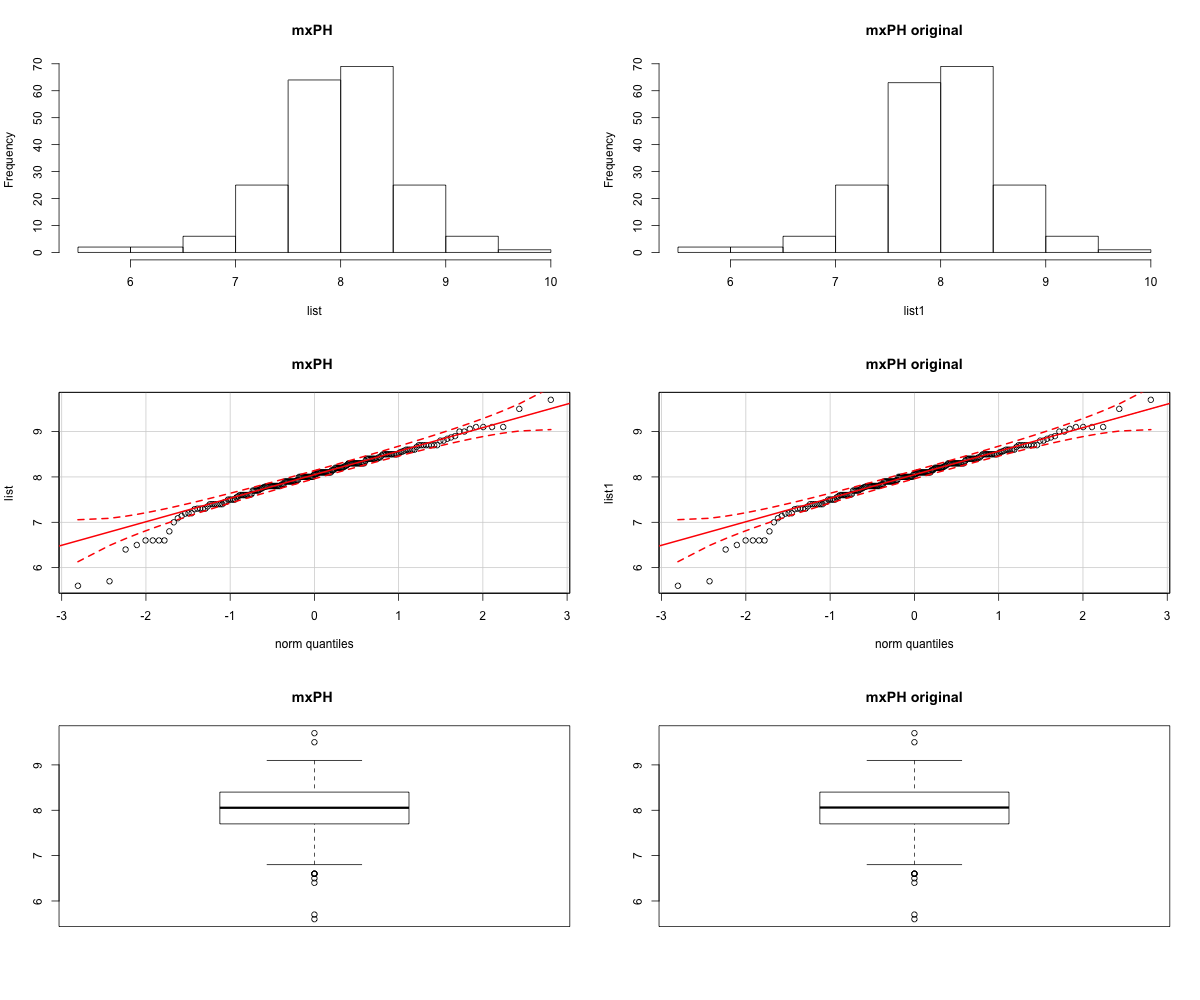


图4通过数据对象之间的相似性来填补缺失值对比图

可以看出，数据被很好的还原，保存了其主要特性。

3 对于7种海藻，分别绘制其数量与标称变量，如size的条件盒图

以a1为例：

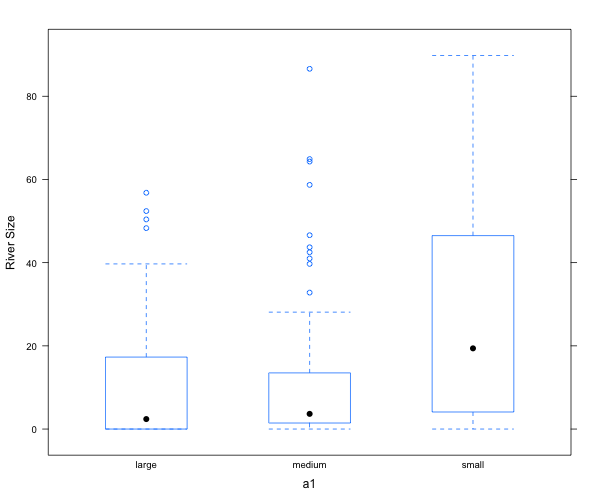


图5 剔除缺失值之后条件盒图

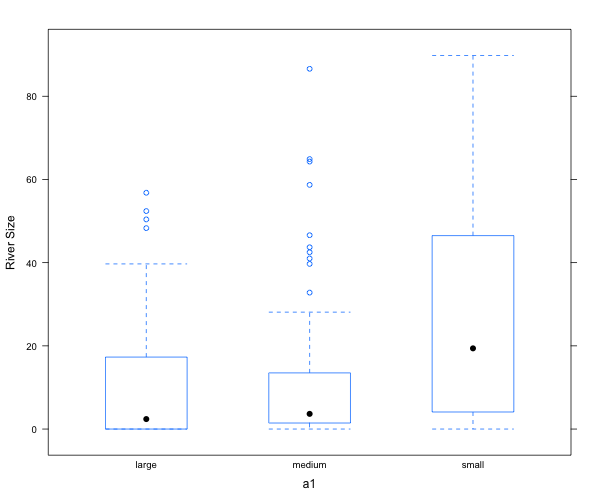


图6用最高频率值来填补缺失值之后条件盒图

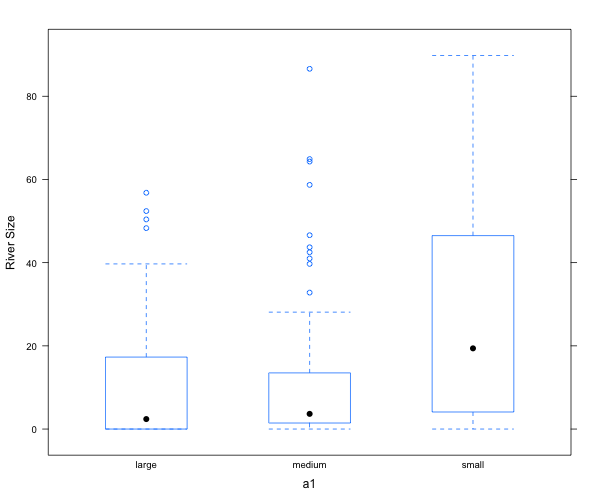


图7属性的相关关系来填补缺失值之后条件盒图

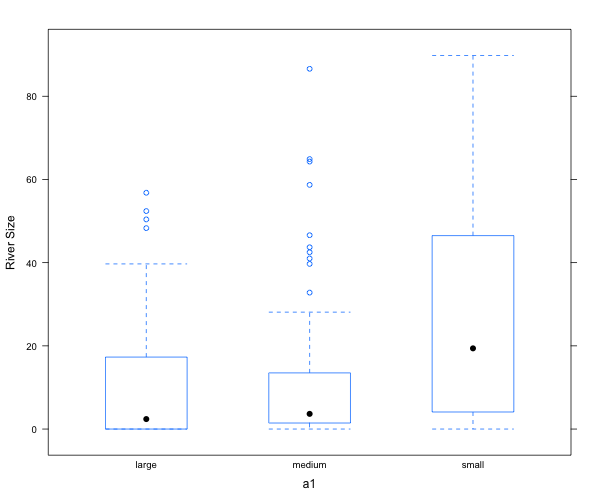


图8通过数据对象之间的相似性来填补缺失值之后条件盒图

4 预处理之后数据

预处理之后的数据为每个对应文件夹下的csv文件