###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。

`original figures`是原始数据的图表。

`fixed figures`是填充完缺失数据之后图表，从图形上来看和原始数据区别不大。

1. 数据摘要

代码：需要导入库DMwR

###################################################

### Loading the Data into R

###################################################

library(DMwR)

#head(algae)

setwd("~/workspace/r-base")

algae <- read.table('Analysis.txt',

header=F,

dec='.',

col.names=c('season','size','speed','mxPH','mnO2','Cl',

'NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4',

'a5','a6','a7'),

na.strings=c('XXXXXXX'))

colum <- list('season','size','speed','mxPH','mnO2','Cl',

'NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4',

'a5','a6','a7')

###################################################

### Data Visualization and Summarization

###################################################

sink("Summary.txt") #output into summary.txt

summary(algae)

sink()

## 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

代码：需要导入库car

###################################################

### Data Visualization

###################################################

library(car)

for (i in 4:18) {

a <- colum[[i]]

png(file=paste('Histogram of ',a,'.png') )

#par(mfrow=c(1,2))

hist(algae[,i], prob=T, xlab='',main=paste('Histogram of ',a))

lines(density(algae[,i],na.rm=T))

rug(jitter(algae[,i]))

dev.off( )

png(file=paste('QQ of ',a,'.png') )

qqPlot(algae[,i],main=paste('Normal QQ plot of ',a),ylab= as.character(a))

dev.off( )

png(file=paste('box of ',a,'.png') )

#par(mfrow=c(1,1))

boxplot(algae[,i],ylab= as.character(a))

rug(jitter(algae[,i]),side=2)

abline(h=mean(algae[,i],na.rm=T),lty=2)

dev.off( )

#identify(algae[,i])

}

library(lattice)

for (i in 12:18) {

a <- colum[[i]]

png(file=paste('Season and ',a,'.png') )

print(bwplot(season ~ algae[,i], data=algae,ylab='Season',xlab=paste('Algal ',a)))

dev.off( )

png(file=paste('River Size and ',a,'.png') )

print(bwplot(size ~ algae[,i], data=algae,ylab='River Size',xlab=paste('Algal ',a)))

dev.off( )

png(file=paste('River Speed and ',a,'.png') )

print(bwplot(speed ~ algae[,i], data=algae,ylab='River Speed',xlab=paste('Algal ',a)))

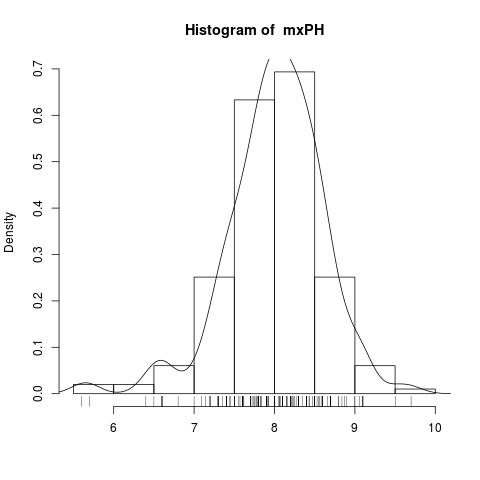
dev.off( )

}

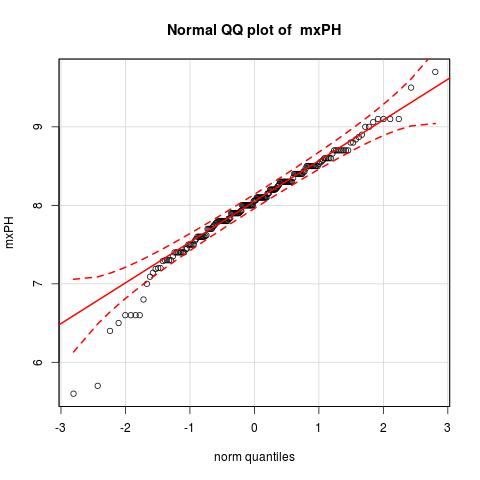
2.1 数值数据

只举例部分属性，其他图片参见仓库。

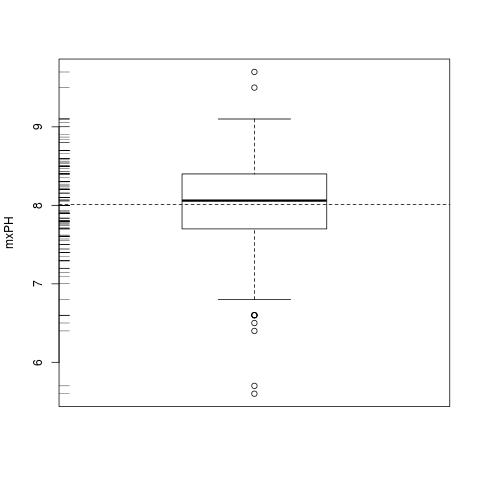
绘制直方图实例：mxPH



绘制Q-Q图实例：mxPH

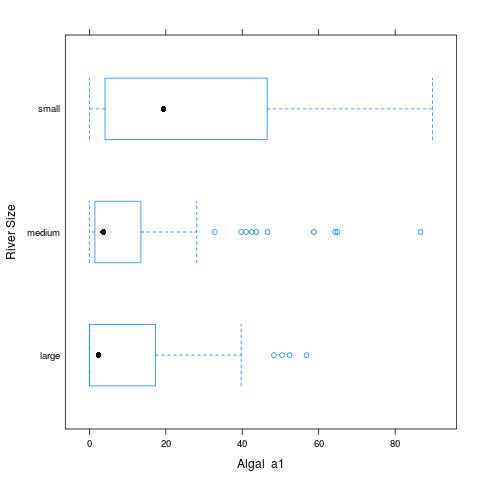


绘制盒图实例：mxPH，同时检测出离群点。

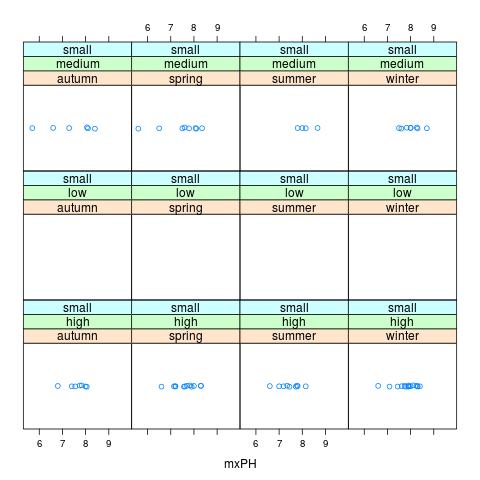


2.2 针对海藻

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

代码：需要导入库lattice

###################################################

### Unkwnon Values Process

###################################################

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

###0,orginal

write.csv(algae,file = "Analysis\_Orginal.csv",na = "XXXXXXX")

###1,Delete

algae <- na.omit(algae)

write.csv(algae,file = "Analysis\_Delete.csv",na = "XXXXXXX")

###2,most Frequency

algae <- read.table('Analysis.txt',

header=F,

dec='.',

col.names=c('season','size','speed','mxPH','mnO2','Cl',

'NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4',

'a5','a6','a7'),

na.strings=c('XXXXXXX'))

for (i in 4:11) {

a <- colum[[i]]

as.numeric(names(table(algae[,as.character(a)])))[which.max(table(algae[,as.character(a)]))]

algae[is.na(algae[,i]),as.character(a)] <- as.numeric(names(table(algae[,as.character(a)])))[which.max(table(algae[,as.character(a)]))]

}

write.csv(algae,file = "Analysis\_Frequency.csv",na = "XXXXXXX")

###3,Correlation

algae <- read.table('Analysis.txt',

header=F,

dec='.',

col.names=c('season','size','speed','mxPH','mnO2','Cl',

'NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4',

'a5','a6','a7'),

na.strings=c('XXXXXXX'))

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

sink("CorrelationMatrix.txt") #output into CorrelationMatrix.txt

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

sink()

sink("CorrelationCoefficient.txt") #output into CorrelationCoefficient.txt

lm(PO4 ~ oPO4,data=algae)

lm(NH4 ~ NO3,data=algae)

lm(Chla ~ mxPH,data=algae)

lm(mxPH ~Chla ,data=algae)

lm(mnO2 ~oPO4 ,data=algae)

lm(Cl ~oPO4 ,data=algae)

sink()

png(file=paste('season-speed-size of mnO2','.png') )

stripplot(~mnO2|season \*speed \*size,data=algae,jitter=T)

dev.off()

png(file=paste('season-speed-size of Cl','.png') )

stripplot(~Cl|season \*speed \*size,data=algae,jitter=T)

dev.off()

png(file=paste('season-speed-size of Chla','.png') )

stripplot(~Chla|season \*speed \*size,data=algae,jitter=T)

dev.off()

png(file=paste('season-speed-size of mxPH','.png') )

stripplot(~mxPH|season \*speed \*size,data=algae,jitter=T)

dev.off()

fillPO4 <- function(oP) {

if (is.na(oP)) return(NA)

else return(42.897 + 1.293 \* oP)

}

algae[is.na(algae$PO4),'PO4'] <- sapply(algae[is.na(algae$PO4),'oPO4'],fillPO4)

fillChla <- function(mP) {

if (is.na(mP)) return(NA)

else return(abs(-139.4 + 19 \* mP))

}

algae[is.na(algae$Chla),'Chla'] <- sapply(algae[is.na(algae$Chla),'mxPH'],fillChla)

fillmxPH <- function(Ch) {

if (is.na(Ch)) return(NA)

else return(7.92896 + 0.01047 \* Ch)

}

algae[is.na(algae$mxPH),'mxPH'] <- sapply(algae[is.na(algae$mxPH),'Chla'],fillmxPH)

fillmnO2 <- function(oP) {

if (is.na(oP)) return(NA)

else return(9.93341 -0.01093 \* oP)

}

algae[is.na(algae$mnO2),'mnO2'] <- sapply(algae[is.na(algae$mnO2),'oPO4'],fillmnO2)

fillCl <- function(oP) {

if (is.na(oP)) return(NA)

else return(28.4771 + 0.1992 \* oP)

}

algae[is.na(algae$Cl),'Cl'] <- sapply(algae[is.na(algae$Cl),'oPO4'],fillCl)

#algae[is.na(algae$mnO2),"mnO2"] <- mean(algae$mnO2,na.rm = T)

write.csv(algae,file = "Analysis\_Correlation.csv",na = "XXXXXXX")

###4,Similarity

algae <- read.table('Analysis.txt',

header=F,

dec='.',

col.names=c('season','size','speed','mxPH','mnO2','Cl',

'NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4',

'a5','a6','a7'),

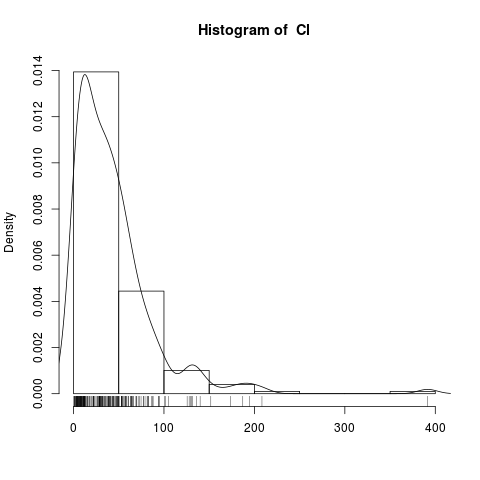
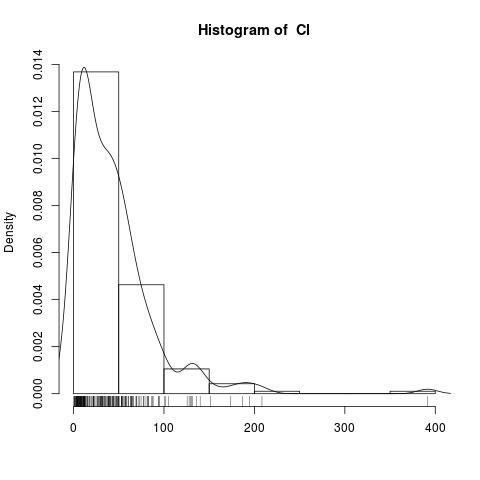
na.strings=c('XXXXXXX'))

algae <- knnImputation(algae,k=10)

write.csv(algae,file = "Analysis\_Similarity.csv",na = "XXXXXXX")

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。