**海藻数据的分析**

2120151046 魏思杰

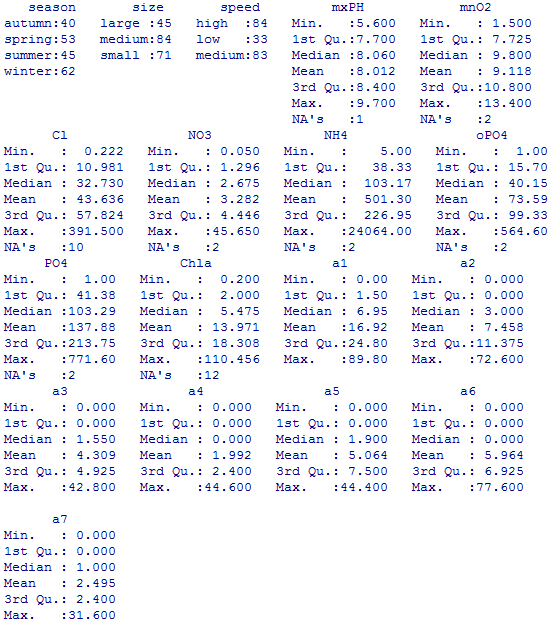
**基于R语言进行数据分析。首先打开Analysis.txt文件，读取表格并加上表头，识别“XXXXXXX”为缺失值：**

Records<-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"))

**得到数据的基本统计量：**

summary(Records)

结果如下：



**统计缺失值个数：**

missing\_mxPH<-nrow(Records[!complete.cases(Records$mxPH),])

missing\_mnO2<-nrow(Records[!complete.cases(Records$mnO2),])

missing\_Cl<-nrow(Records[!complete.cases(Records$Cl),])

missing\_NO3<-nrow(Records[!complete.cases(Records$NO3),])

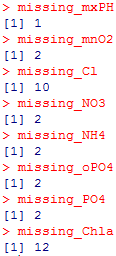
missing\_NH4<-nrow(Records[!complete.cases(Records$NH4),])

missing\_oPO4<-nrow(Records[!complete.cases(Records$oPO4),])

missing\_PO4<-nrow(Records[!complete.cases(Records$PO4),])

missing\_Chla<-nrow(Records[!complete.cases(Records$Chla),])

结果：



**画出每个数值属性的直方图，并储存为png文件：**

png(file="hist\_mxPH.png", bg="transparent")

hist(Records$mxPH, prob=T, main="hist plot:mxPH")

dev.off()

png(file="hist\_mnO2.png", bg="transparent")

hist(Records$mnO2, prob=T, main="hist plot:mnO2")

dev.off()

png(file="hist\_Cl.png", bg="transparent")

hist(Records$Cl, prob=T, main="hist plot:Cl")

dev.off()

png(file="hist\_NO3.png", bg="transparent")

hist(Records$NO3, prob=T, main="hist plot:NO3")

dev.off()

png(file="hist\_NH4.png", bg="transparent")

hist(Records$NH4, prob=T, main="hist plot:NH4")

dev.off()

png(file="hist\_oPO4.png", bg="transparent")

hist(Records$oPO4, prob=T, main="hist plot:oPO4")

dev.off()

png(file="hist\_PO4.png", bg="transparent")

hist(Records$PO4, prob=T, main="hist plot:PO4")

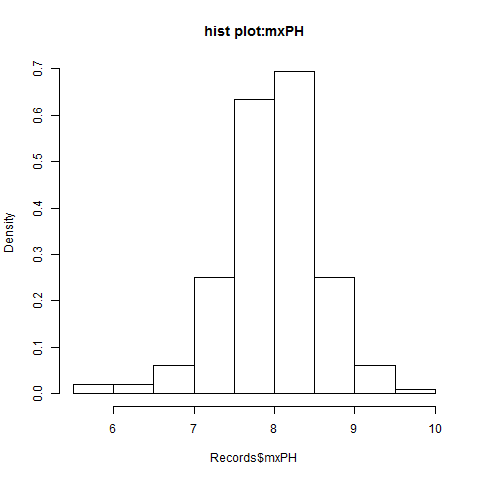
dev.off()

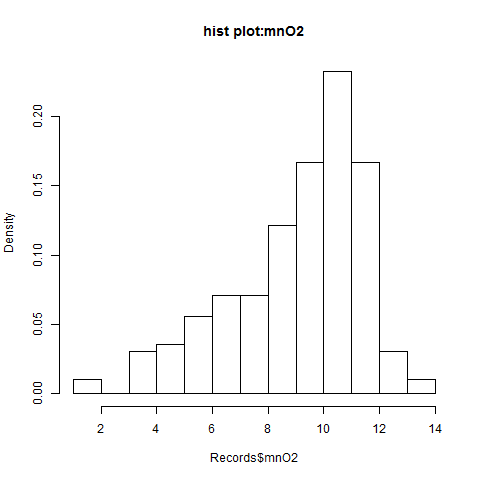
png(file="hist\_Chla.png", bg="transparent")

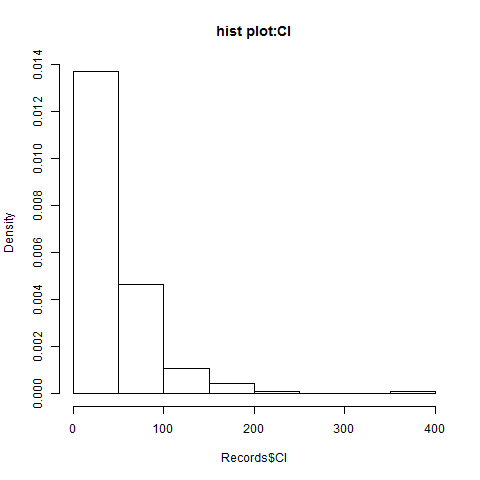
hist(Records$Chla, prob=T, main="hist plot:Chla")

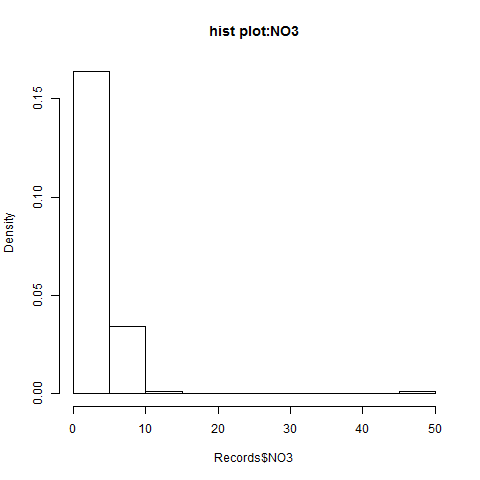
dev.off()

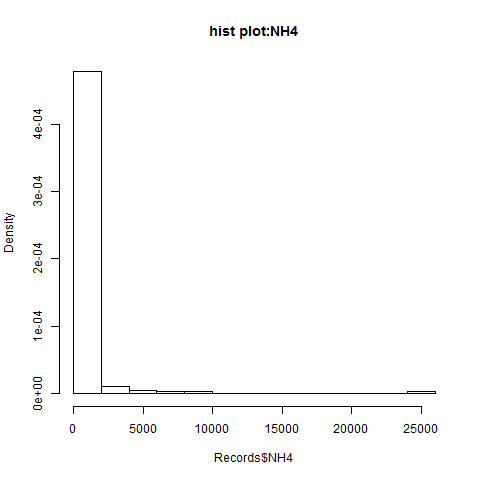
结果：

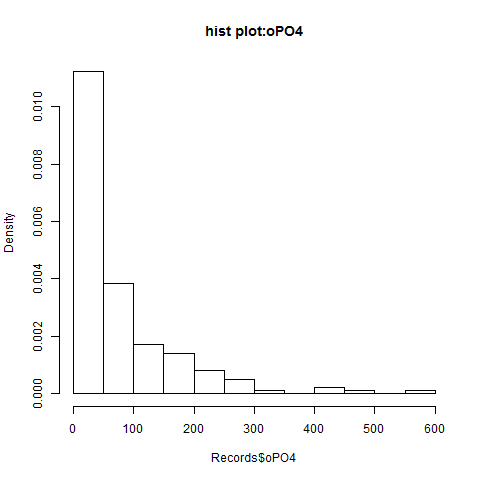


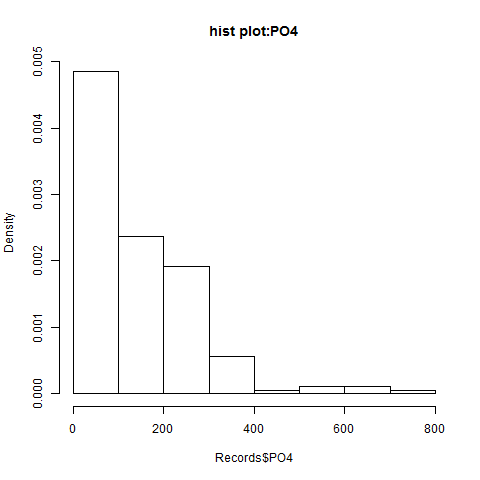


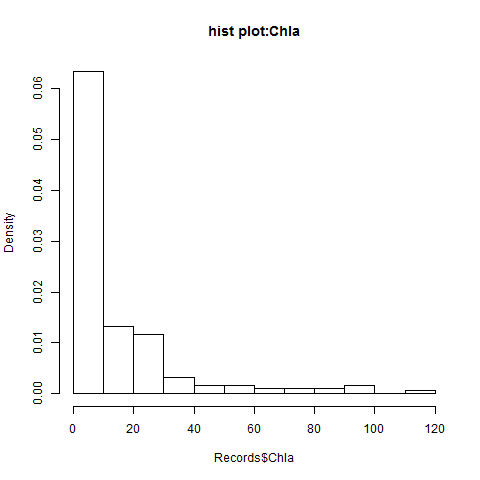












**画出每个数值属性的Q-Q图，并储存为png文件：**

png(file="qq\_mxPH.png", bg="transparent")

qqnorm(Records$mxPH, main="Q-Q plot:mxPH")

qqline(Records$mxPH)

dev.off()

png(file="qq\_mnO2.png", bg="transparent")

qqnorm(Records$mnO2, main="Q-Q plot:mnO2")

qqline(Records$mnO2)

dev.off()

png(file="qq\_Cl.png", bg="transparent")

qqnorm(Records$Cl, main="Q-Q plot:Cl")

qqline(Records$Cl)

dev.off()

png(file="qq\_NO3.png", bg="transparent")

qqnorm(Records$NO3, main="Q-Q plot:NO3")

qqline(Records$NO3)

dev.off()

png(file="qq\_NH4.png", bg="transparent")

qqnorm(Records$NH4, main="Q-Q plot:NH4")

qqline(Records$NH4)

dev.off()

png(file="qq\_oPO4.png", bg="transparent")

qqnorm(Records$oPO4, main="Q-Q plot:oPO4")

qqline(Records$oPO4)

dev.off()

png(file="qq\_PO4.png", bg="transparent")

qqnorm(Records$PO4, main="Q-Q plot:PO4")

qqline(Records$PO4)

dev.off()

png(file="qq\_Chla.png", bg="transparent")

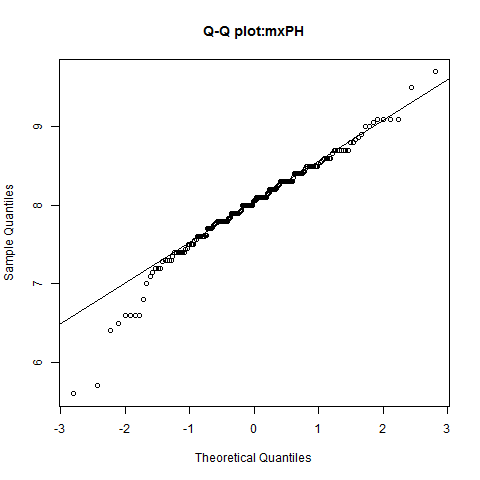
qqnorm(Records$Chla, main="Q-Q plot:Chla")

qqline(Records$Chla)

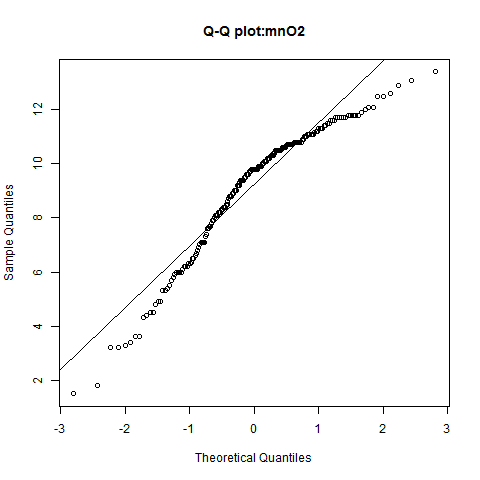
dev.off()

利用Q-Q图鉴别样本数据是否近似于正态分布，只需看QQ图上的点是否近似地在一条直线附近，且该直线的斜率为标准差，截距为均值。

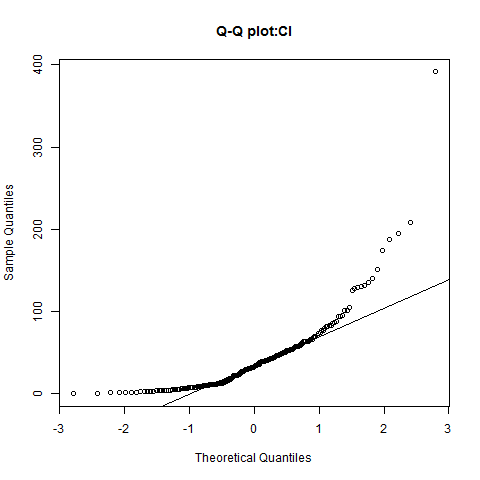
结果：



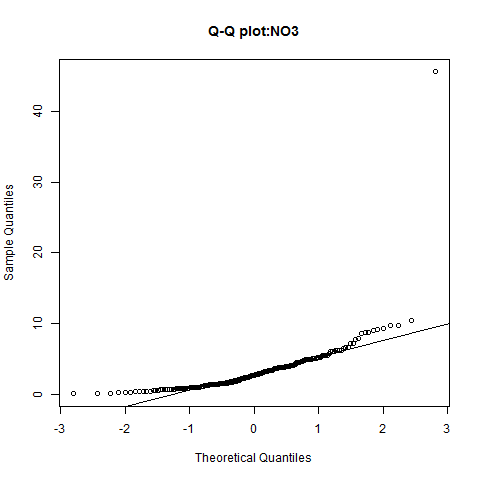
数据呈现为正态分布。



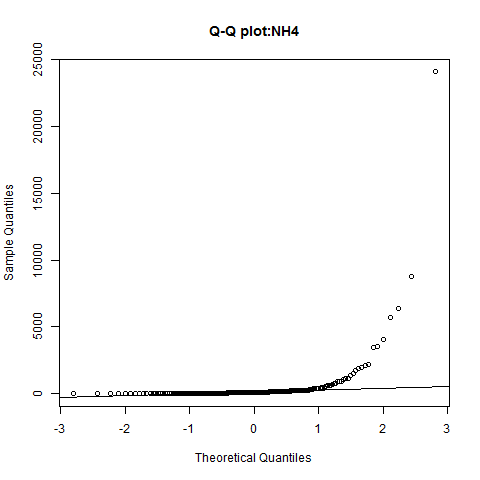
数据不呈现为正态分布。



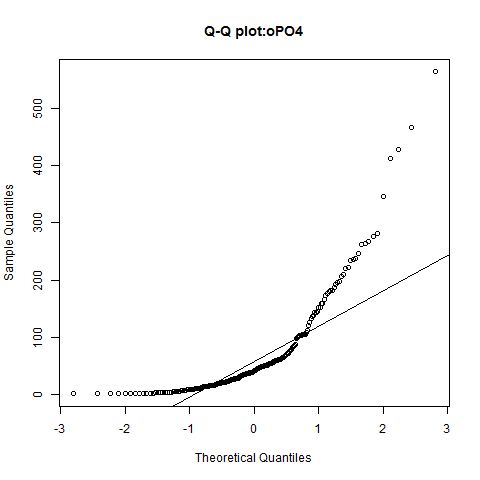
数据不呈现为正态分布。



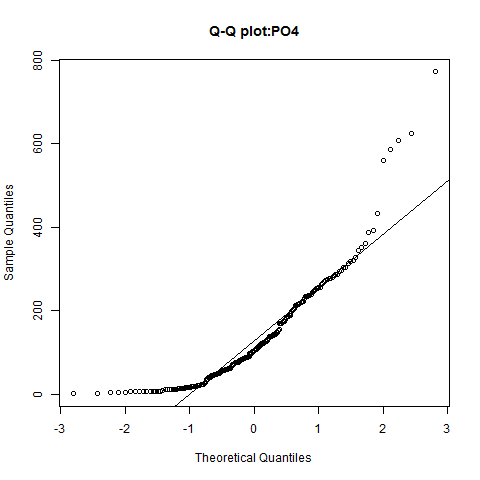
数据呈现为正态分布。



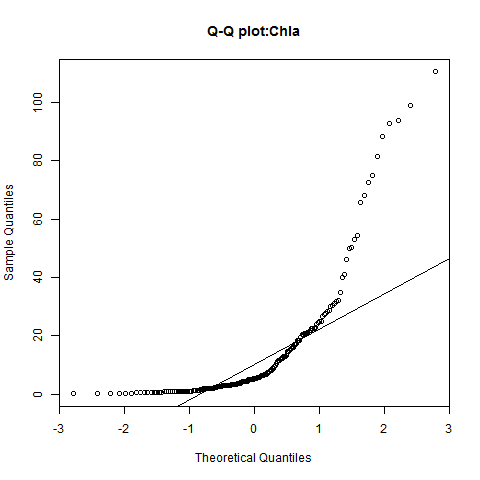
数据可以认为呈现为正态分布。



数据不呈现为正态分布。



数据不呈现为正态分布。



数据不呈现为正态分布。

**画出每个数值属性的盒图，并储存为png文件：**

png(file="box\_mxPH.png", bg="transparent")

boxplot(Records$mxPH,col=c("steelblue"), main="box plot:mxPH")

dev.off()

png(file="box\_mnO2.png", bg="transparent")

boxplot(Records$mnO2,col=c("mediumturquoise"), main="box plot:mnO2")

dev.off()

png(file="box\_Cl.png", bg="transparent")

boxplot(Records$Cl,col=c("sandybrown"), main="box plot:Cl")

dev.off()

png(file="box\_NO3.png", bg="transparent")

boxplot(Records$NO3,col=c("hotpink"), main="box plot:NO3")

dev.off()

png(file="box\_NH4.png", bg="transparent")

boxplot(Records$NH4,col=c("red"), main="box plot:NH4")

dev.off()

png(file="box\_oPO4.png", bg="transparent")

boxplot(Records$oPO4,col=c("yellow"), main="box plot:oPO4")

dev.off()

png(file="box\_PO4.png", bg="transparent")

boxplot(Records$PO4,col=c("green"), main="box plot:PO4")

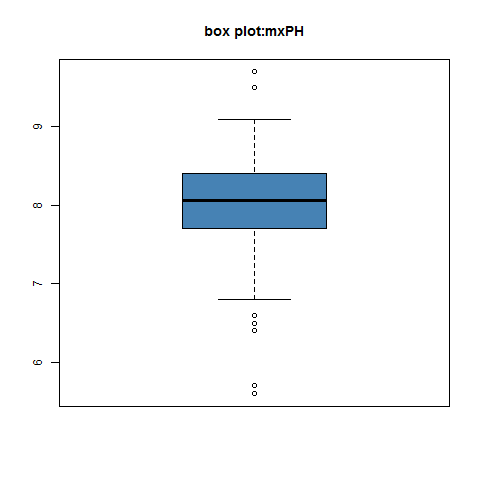
dev.off()

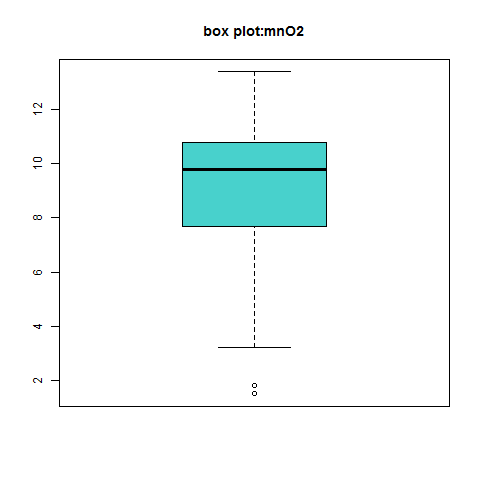
png(file="box\_Chla.png", bg="transparent")

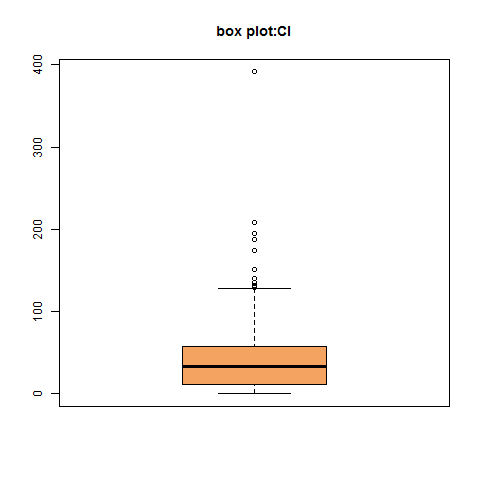
boxplot(Records$Chla,col=c("orange"), main="box plot:Chla")

dev.off()

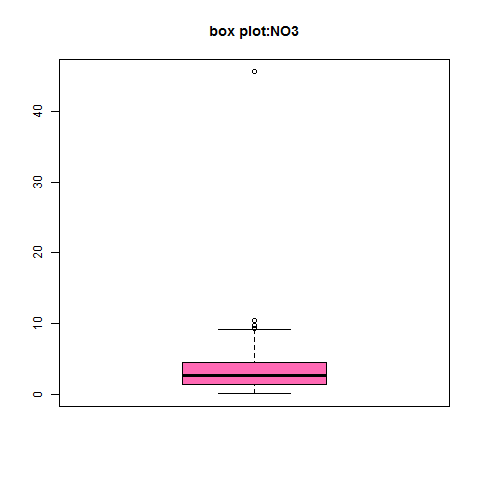
结果：



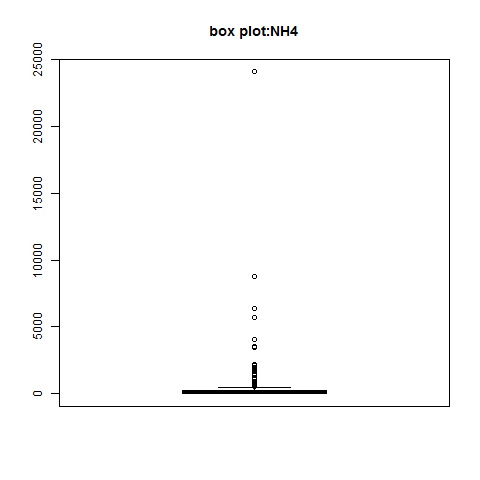




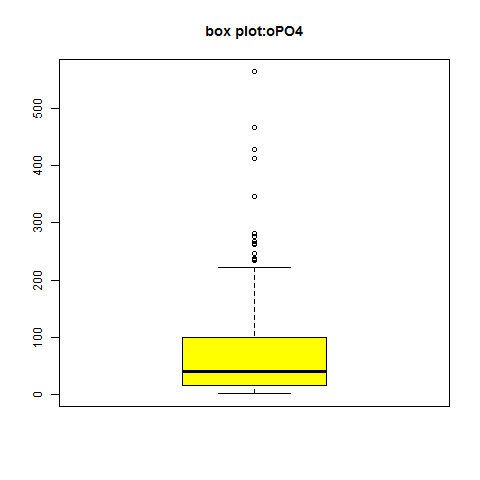
最上方点严重异常，为第六列第153个数据。

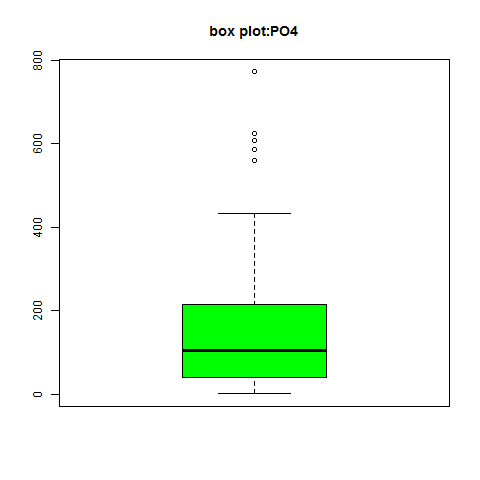


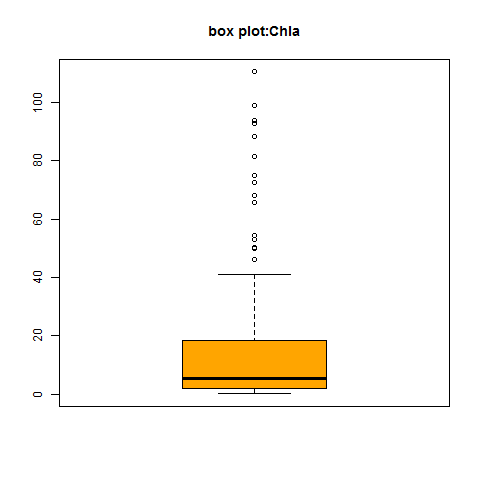
最上方点严重异常，为第七列第153个数据。



最上方点严重异常，为第八列第153个数据。







**数据缺失的处理**

将缺失部分剔除，并将所得数据保存为Analysis\_1.txt

Records\_1<-na.omit(Records)

write.table(Records\_1, file='Analysis\_1.txt', row.names=F, quote=F)

用最高频率值来填补缺失值，并将所得数据保存为Analysis\_2.txt

library(DMwR)

data(Records)

Records\_2<-Records[-manyNAs(Records),]

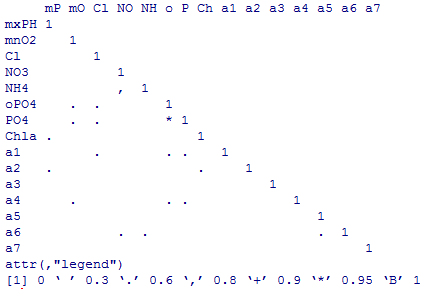
Records\_2<-centralImputation(Records)

write.table(Records\_2, file='Analysis\_2.txt', row.names=F, quote=F)

由于缺乏相关知识，所以难以通过属性的相关关系来填补缺失值，只能分析属性的相关关系：

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

结果：



通过数据对象之间的相似性来填补缺失值，并将所得数据保存为Analysis\_3.txt

#Records\_3<-Records[-manyNAs(Records),]

Records\_3<-knnImputation(Records\_2, k=10)

write.table(Records\_3, file='Analysis\_3.txt', row.names=F, quote=F)