**分析报告**

学院：计算机学院

姓名：王子硕

学号：2120151044

1. **数据摘要**

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

使用R代码：

mydata<- 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',

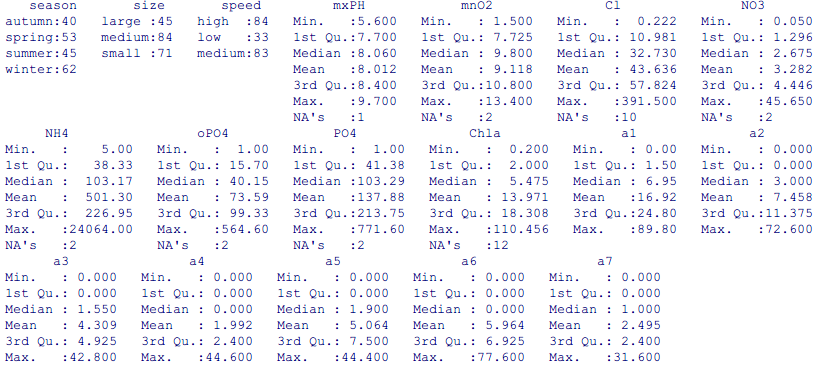
'a5','a6','a7'),

na.strings=c('XXXXXXX'))

用Summary函数分析数据摘要：

summary(mydata)

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：



1. **数据的可视化**
2. 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。

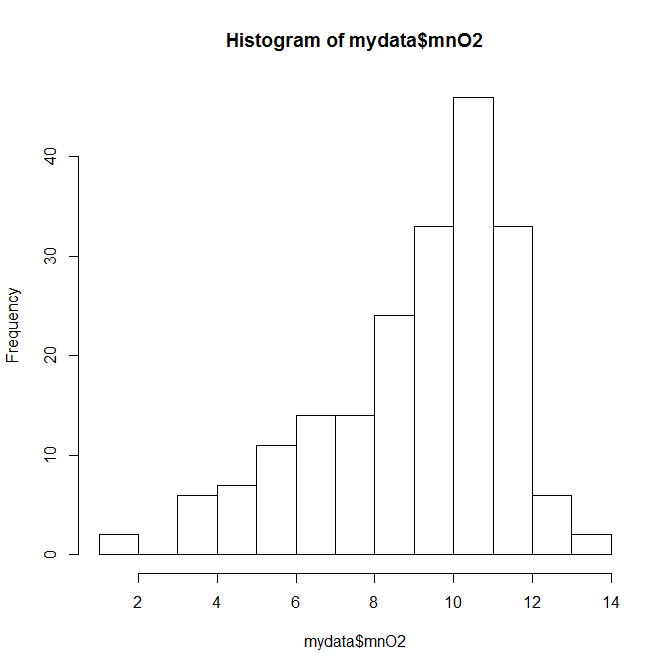
hist(mydata$mxPH)

library(car)

qqPlot(mydata$mxPH,main='Norm QQ Plot of mxPH')



mxPH的直方图与QQ图



mnO2的直方图与QQ图



Cl的直方图与QQ图



NO3的直方图与QQ图



NH4的直方图与QQ图



oPO4的直方图与QQ图



PO4的直方图与QQ图



Chla的直方图与QQ图

1. 绘制盒图，识别离群点

执行如下R代码：

boxplot(mydata$mxPH,ylab='mxPH')

rug(mydata$mxPH,side=4)

abline(h=mean(mydata$mxPH,na.rm=T),lty=2)

Rug函数绘制了每个点在纵轴上的投影情况，abline则绘制了数据的均值，在图中以虚线的方式呈现。由各属性的盒图，可以分析出离群点的数量以及分布情况。



mxPH盒图mnO2盒图



Cl盒图 NO3盒图



NH4 oPO4



PO4盒图 Chla盒图

1. 对七种海藻，绘制其数量与河流大小的条件盒图

使用R代码，绘制其与河流大小的条件盒图，命令如下：

library(lattice)

bwplot(size~a1,data=mydata,ylab='River Size',xlab='a1')

下图反映a1海藻在不同河流大小条件下的盒图形状。依次绘制a1-a7海藻的条件盒图，如图



a1海藻与河流大小的条件盒图 a2海藻与河流大小的条件盒图



a3海藻与河流大小的条件盒图 a4海藻与河流大小的条件盒图



a5海藻与河流大小的条件盒图 a6海藻与河流大小的条件盒图



a7海藻与河流大小的条件盒图

从图中可以看出a1有更高的频数，但是 a3,,a5,a6在中型河流中更多一些。

1. **数据缺失的处理**
2. 将缺失部分剔除

剔除缺失数据与写入文件的命令如下：

omitdata = na.omit(mydata) 剔出缺失数据

write.table(omitdata,'OmitedData.txt',col.names = F,row.names = F, quote = F)写入文件

1. 使用高频数值来填补缺失值

library(DMwR)

preprocess2 = mydata[-manyNAs(mydata),]

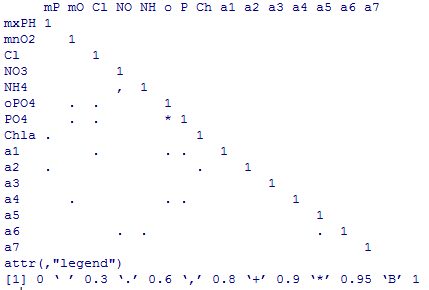
preprocess2 = centralImputation(Preprocess2)

write.table(preprocess2,'D:/DataMining/CentralImputationData.txt',col.names = F,row.names = F, quote = F)

1. 通过属性的相关关系来填补缺失值

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

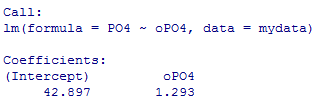
得到属性之间的相关性如下图：



从图中可以看出oPO4与PO4相关度超过0.9，所以可以用这两个属性作相关分析，互相填补缺失数据。

使用一下代码得到其线性模型：

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



oPO4与PO4的线性模型分析

得到结果如图28所示，表示得到的线性模型为PO4=42.897 + oPO4\* 1.293.

使用线性模型来填充PO4与oPO4的数据：

preprocess3 = mydata[-manyNAs(mydata),]

fillPO4 <- function(oP){

if(is.na(oP))

return(NA)

else return (42.897 + 1.293 \* oP)

}

preprocess3[is.na(preprocess3$PO4),'PO4'] = sapply(preprocess3[is.na(preprocess3$PO4),'oPO4'],fillPO4)

1. 使用数据对象之间的相似型填补缺失值

preprocess4 = knnImputation(mydata,k=10)

write.table(preprocess4,'D:/DataMining/knnImputationData.txt',col.names = F,row.names = F, quote = F)