**Data Mining Homework**

**kwon dayeon**

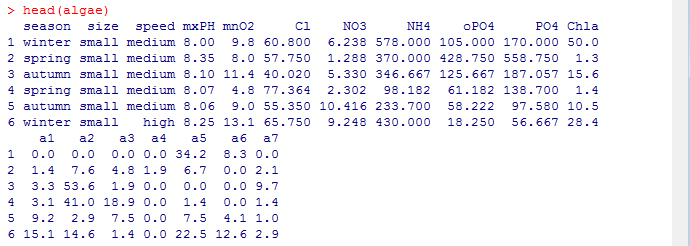
**2820150089**

The experiment of data processing and visualization display by using R language

## 1.Data Reading

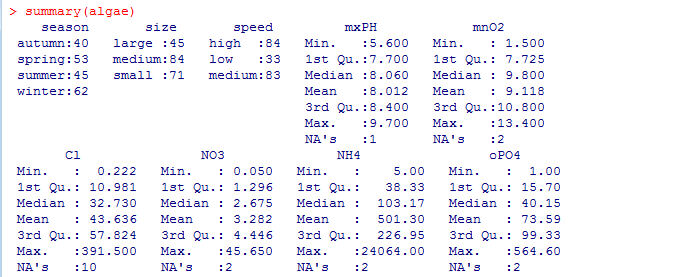
algae <- read.table(file = 'Analysis.txt', col.names = c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'), na.strings = c('XXXXXXX'))

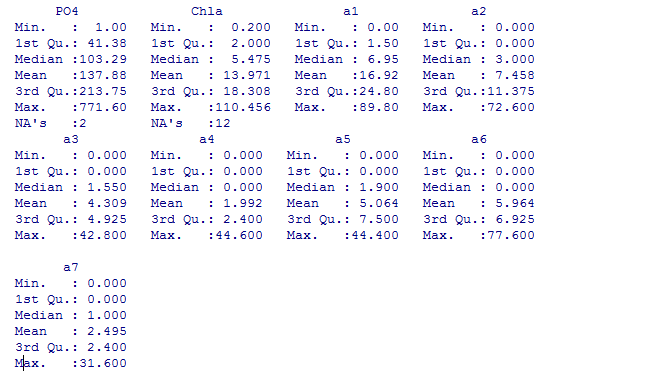
Read Successful



**2.Summary of Data and visualization**

**2.1 data summary**

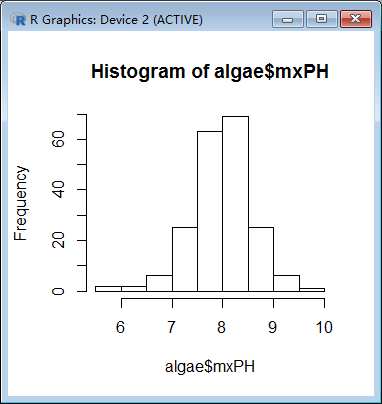




This simple function gives the data statistics describing their characteristics, and his name variable, he gives each of the values of the frequency of the variables.For example, the results from the war in the winter acquisition of Blockage of the size is the largest rivers, more than any other season with rivers the flow rate of the low, 45 fewer.For value type variables, R provides us a quarter of a man, median, four quarters of the three digits, value, are a series of information.The statistical information provides variable value the distribution of the variables include impairment of value in the preliminary information, the values of the cases, the string behind the DNA for impairment of a number, through the mean and median, quartile of the information.We can understand the data distribution of degrees and diversification.And these information graphics through most of them can be expressed.

### 2.2 data visualization

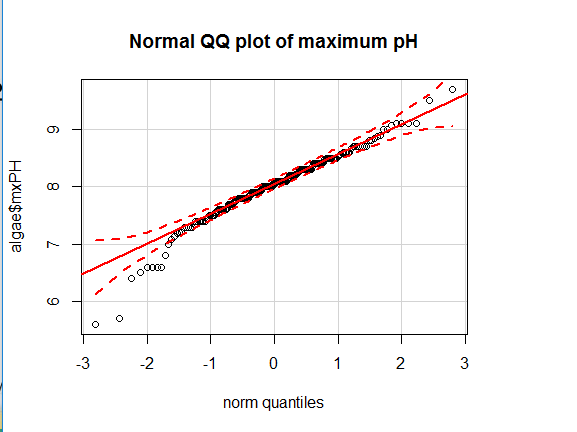
（1）drawing histogram：mxPH

hist(algae$mxPH)

（2）Draw Q-Q diagrams：mxPH

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

> qq.plot(algae$mxPH,main="Normal QQ plot of maximum pH")



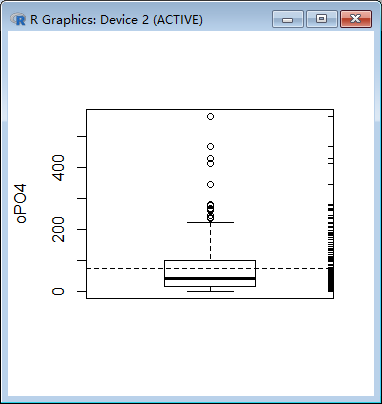
Draws normal distribution and variables in quartile positively correlated the theory of the nodes Fig., at the same time, given the number of normal distribution of confidence the confidence of the area of strip between the map, from the map there are several variables, little apparent value of the 95 % The confidence interval. They don't obey the law of normal distribution.

（3）draw box diagrams：oPO4

> boxplot(algae$oPO4,ylab='oPO4')

> rug(algae$oPO4,side=4)

> abline(h=mean(algae$oPO4,na.rm=T),lty=2)



Boxes above the small horizontal line above the small circles represent particularly large compared with other values of value, usually considered to be outliers. The distribution of the variable oPO4 is concentrated in the smaller range of the observed value, so the distribution is positive. Most water samples of the oPO4 are relatively low, there are several water samples of the higher, and even particularly high.

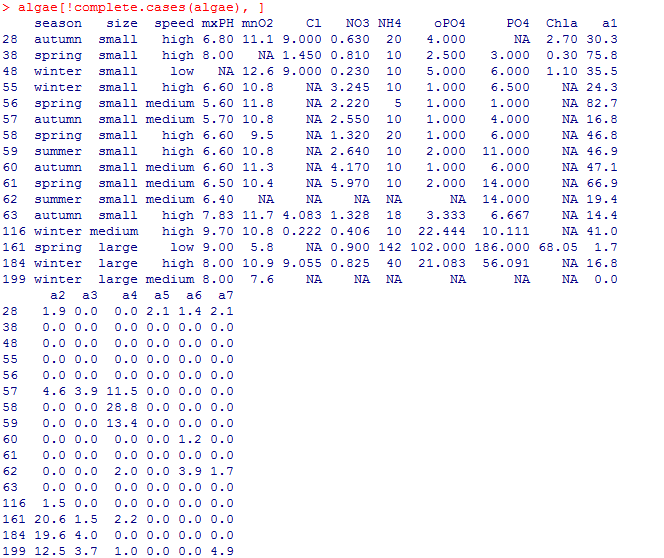
## The treatment of missing data

## 3.1 Excluding missing values

Excluded records with missing data is easy to do, especially when the proportion of those records available when the data set is small, this choice is reasonable.

1.Check the records with missing values

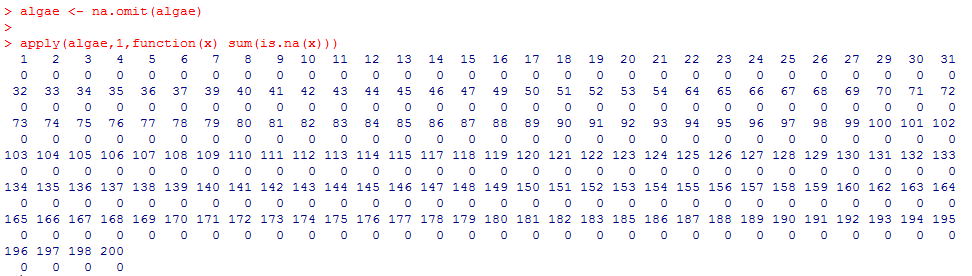
> algae[!complete.cases(algae), ]



2. Excluding missing values

> algae <- na.omit(algae)

> apply(algae,1,function(x) sum(is.na(x)))



3.Save the preprocessed data set

> write.table(algae,file = 'Analysis\_1.txt',quote=FALSE,sep='\t',row.names =FALSE, col.names = FALSE)

### 3.2 fill shortage of value with the highest frequency loss of value

Fill a record value of another method is to try to find the missing value

the value of the most likely.

1. Using centralImputation () from trend to use the data center to fill data value of all impairment.

> library(DMwR)

> algae<-centralImputation(algae)

2. Save the preprocessed data set

> write.table(algae,file = 'Analysis\_2.txt',quote=FALSE,sep='\t',row.names =FALSE, col.names = FALSE)

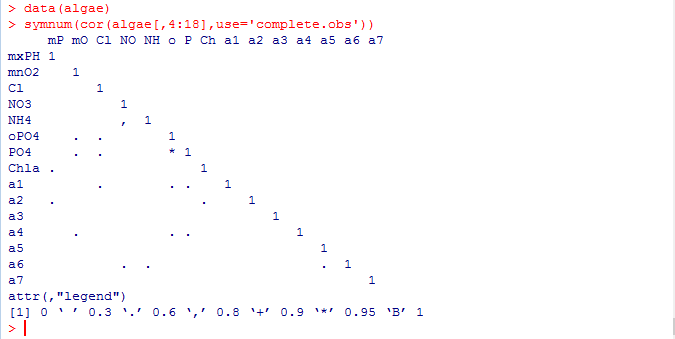
### 3.3 Correlation to fill a shortage of property through lost value

Another acquisition of impairment deviation of less value estimate of the way is to explore the correlation between the variables.For example, through the variable value can find some variables, the correlation between the highly correlated with mxPH.

1. Gets the variables the correlation matrix

> data(algae)

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



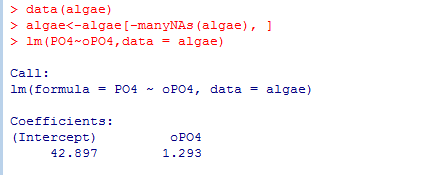
Results showed that most of the variables are not related, however, there are two exceptions: variable between NH4 and NO3, variable between PO4 and oPO4. Between the latter two variables related to a high value (greater than 0.9). And, therefore, identify missing data by them is dangerous. In addition, because 199 62 samples and sample there are too many variables with missing values, so if they are excluded, samples of NH4 and NO3 in the variables without missing values.

2. Look for variable linear relationships between PO4 and oPO4

> data(algae)

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

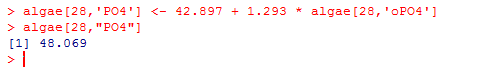
> lm(PO4~oPO4,data = algae)



Linear model is PO4=1.29oPO4+42.90, if these variables are not at the same time has missing values, you can use this formula to calculate the missing values of these variables.

3. After eliminating 199 62 samples and samples ,left a sample28

In PO4 has missing values, you can simply use the above linear relation calculate fill values for missing values



By 28,PO4 correlation between variables obtained water samples to fill the missing value for 48

4. Save the preprocessed data set

> write.table(algae,file = 'Analysis\_3.txt',quote=FALSE,sep='\t',row.names =FALSE, col.names = FALSE)

### 3.4 Filling missing data values through similarities between objects

Correlation between dataset is different from exploration column, try using the similarity between rows to fill in the missing values.

We can use this method to fill out the two contains too much NA samples of other missing data. Methods as described above if the two samples are similar, some of these variables have missing values, the missing data may be related to other water values are similar. In order to use this intuitive method, you first define the concept of similarity. Similarity is often described by multiple metric space defined by the variables of the observations. In the literature there are many metric similarity index used is the Euclidean distance.

This distance can be informal, defined as any two cases sum of the squares of the difference between the observed value. The method described below is to use this measure to find any cases with missing values similar to those of 10 water, and use them to fill in the missing values. We consider two methods of applying these values. The first method is simple to calculate the closest case 10 digits with the digits to fill in these missing values. Similar to the second method is to use the weighted mean of the data. Weight size increases with distance from the outstanding cases of missing values for d, then its value is a weighted average of the weights: w d = ��−��.

1. Above method can add functions of knnImputation () to achieve this function to find any with a variant of the Euclidean distance k-nearest neighbor.

>algae<-read.table(file='Analysis.txt',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'))

> library(DMwR)

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

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

2. Save the processed data

> write.table(algae,file = 'Analysis\_4.txt',quote=FALSE,sep='\t',row.names =FALSE, col.names = FALSE)