**DATA MINING**

**HOMEWORK - 1**

**NAME: SEMHANDO, JOHN PETER**

**ID: 2820150014**

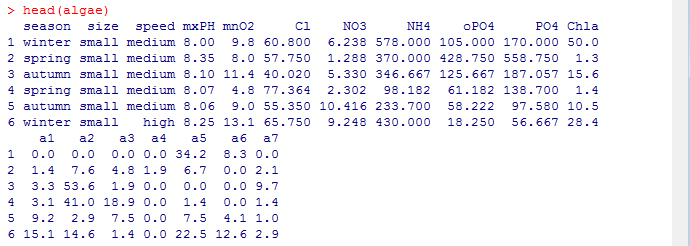
The experiment of data processing and visualization display by using R language

**Data reading**

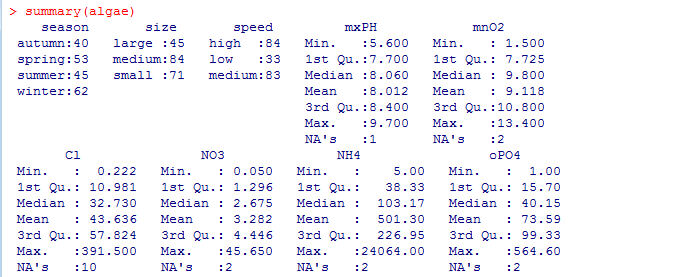
R software is used to load DMWR add package

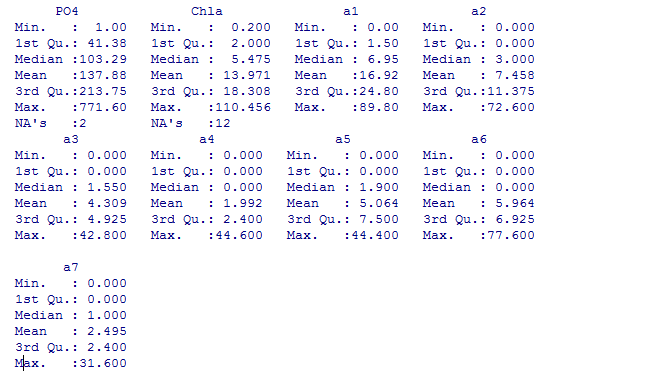
algae <- read.table(file = 'Analysis.txt', col.names = c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4','oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'), na.strings = c('XXXXXXX'))

Read Successful



**Summary of Data and Visualization**



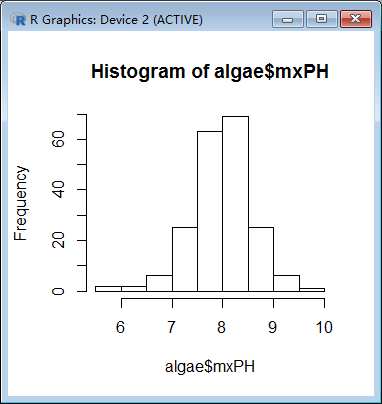


This simple function gives a statistical summary of the data, for nominal variables, he gives the value of each variable frequency. For example, results from samples collected in winter than any other season in more rivers size to a maximum of 45, low river flow rates less. For a numeric variable, r provides us with one-fourth digits, median, mean, three-fourths, extreme and a series of information. These statistics provide the value of a variable distribution of information, in the case of variables have missing values, string value that is later in NA as the number of missing values, through the median, mean, four-digit information, we can understand the skewness of the distribution and dispersion of the data.

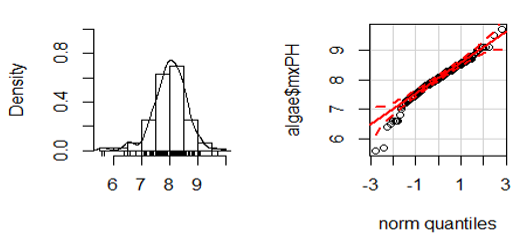
**Visualization**

The above program draws the histogram of the variable mxPH. As shown in the following picture, the distribution of the variable mxPH very close to normal distribution, its value and most have focused on variables near the mean, we do the Q-Q to test whether the variable is a normal distribution, the package is loaded car, the use of qq.plot () drawing Q-Q plot

To draw a histogram: mxPH Hist (algae$mxPH)



To draw Q-Q diagrams: mxPH > par (mfrow=c (1,2)) > qq.plot(algae$mxPH,main="Normal QQ plot of maximum pH")



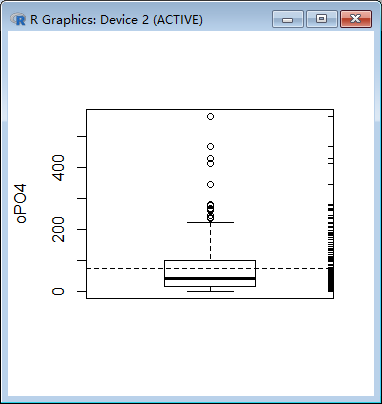
On the left side of the figure is the histogram of the kernel density curve, which can be easily distinguished from the actual value of the variable near the X axis, for example, we can find that there are two words that are significantly lower than all the other values. This kind of data is very important, because it can determine the possible errors in the data sample. The Q-Q chart on the right, he painted the variable value and the theory of normal distribution quantile scatter diagram, also gives his normal distribution of 95% confidence interval of the strip chart, from the right-hand variables have several small value obviously in the 95% confidence interval, they refused to accept from a normal distribution.

**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

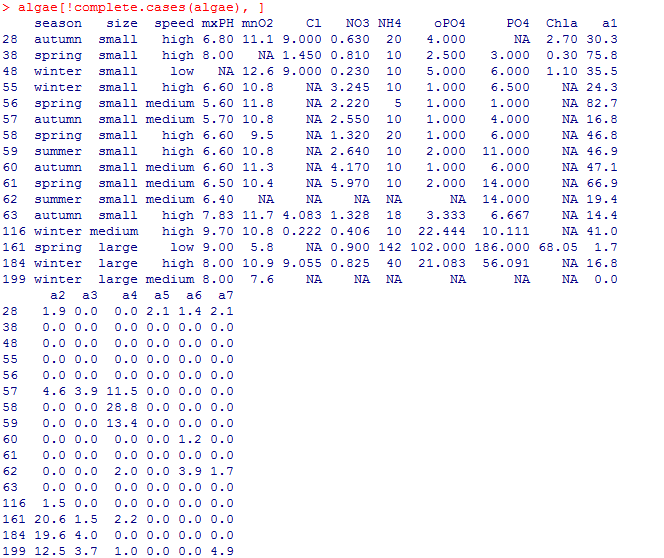
**The treatment of missing data**

**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

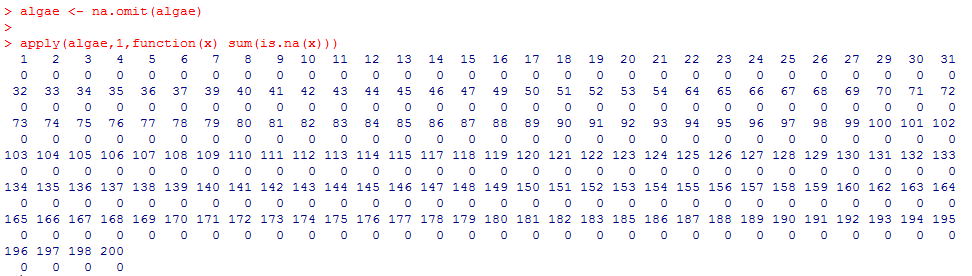
> algae[!complete.cases(algae), ]



1. Excluding missing values

> algae <- na.omit(algae)

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



1. Save the preprocessed data set

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

**The highest frequency values are used to fill in missing values**

Filled with records of missing value another way is to try to find the most probable values for missing values.

1. use central Imputation () to automatically fill all the missing data values with data center trends.

> library(DMwR)

> algae<-centralImputation(algae)

1. Save the preprocessed data set

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

**By the correlation between property to fill in the missing values**

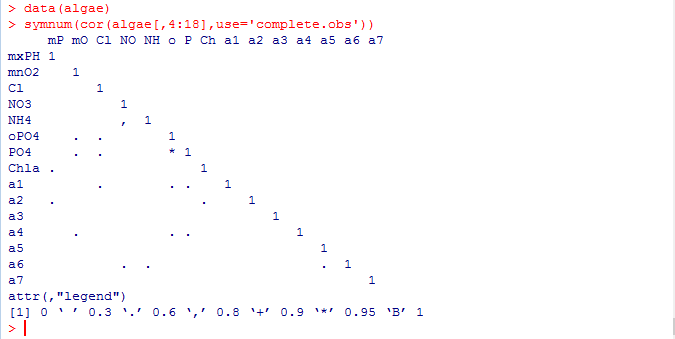
Another gets less missing value estimation method is to explore relationships between variables.

For example, the correlation between the value of the variable, some variables are highly correlated with mxPH can be found.

1. Access to 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. The correlation between variables NH4 and NO3 was not particularly significant (0.72), so it is dangerous to identify missing data based on them. In addition, because the sample 62 and sample 199 have too many variables that contain missing values, so if you remove them, the variables in the sample NH4 and NO3 have no missing values. As for the variables PO4 and oPO4, the correlation between them can help to fill the missing values of these two variables.

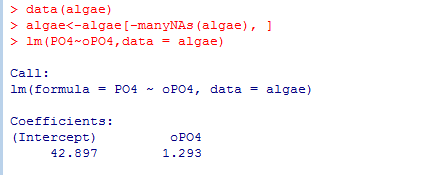
Furthermore, because 199 62 samples and sample there are too many variables with missing values, so if they are excluded, samples of NH4 and NO3 in the variables without missing values.

1. The linear relationships between variables PO4 and oPO4 is as follows

> data(algae)

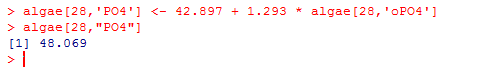
> algae<-algae[-manyNAs(algae), ]

> 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.

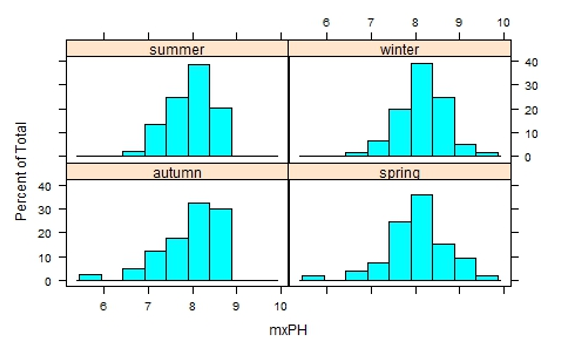
1. After eliminating sample 199 and sample 62 samples, there was one sample of 28 missing in the PO4 values, you can simply use the above linear relation calculate fill values for missing values.



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

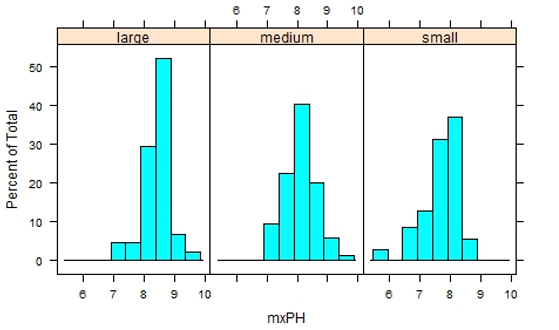
Studies on linear relationships allows us to fill some of the missing values. However, there are several observations with missing values. You can try to explore the case data contains missing values of the relationship between the variables and the name of the variable, which can be added to the R package lattice function to draw the conditional histogram. Drawing code is as follows:

> histogram(~mxPH|season,data=algae)



The above picture is drawn in different seasonal variable mxPH histogram. MxPH histogram in different seasons is not significantly different, the river below the size of size to carry out the above similar analysis, the executive order:

> histogram(~mxPH|season,data=algae)

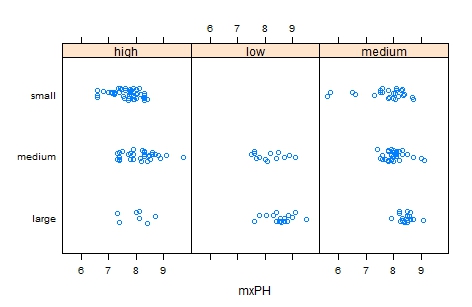


From the obtained histograms, smaller rivers have smaller mxPH values.For this kind of correlation study, it can be extended to more than one nominal variable:

> stripplot(size~mxPH|speed,data=algae,jitter=T)

Instruction above results as shown below, the parameter jitter=T that Y values to a rand-om permutation of the small scope, which can avoid the same value between overlap each other to lead to loss of some specific values of the observed with focus on the value of the information.

This type of analysis can be applied to other variables that contain missing values. However, this analysis is a tedious process, with a large number of variables that need to be analyzed. Analysis of data sets with less nominal variables.

****

1. Save the processed data

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

**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 = 𝑒−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)

In case you use the median to fill in the missing values, you can use the following code

> algae<-knnImputation(algae,k=10,meth = "median")

Basically, data is no longer contain missing values NA in the usage of these simple operation, the use of R to analyze the other functions are fully prepared.

algae[!complete.cases(algae),]

1. Save the preprocessed data set

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