**Data Mining Assignment**

**-1-**

Student ID : 2820150081

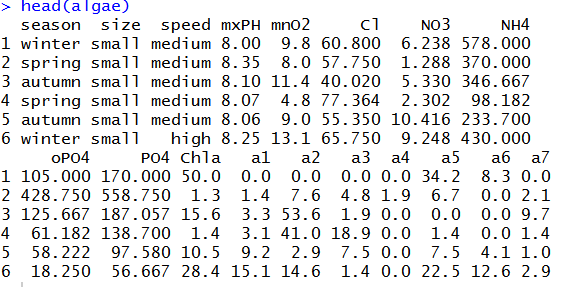
Name : Lee SuJin

The experiments using R language for data processing and visualization display.

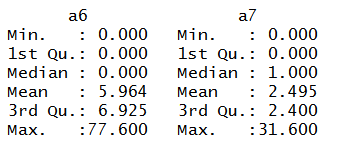
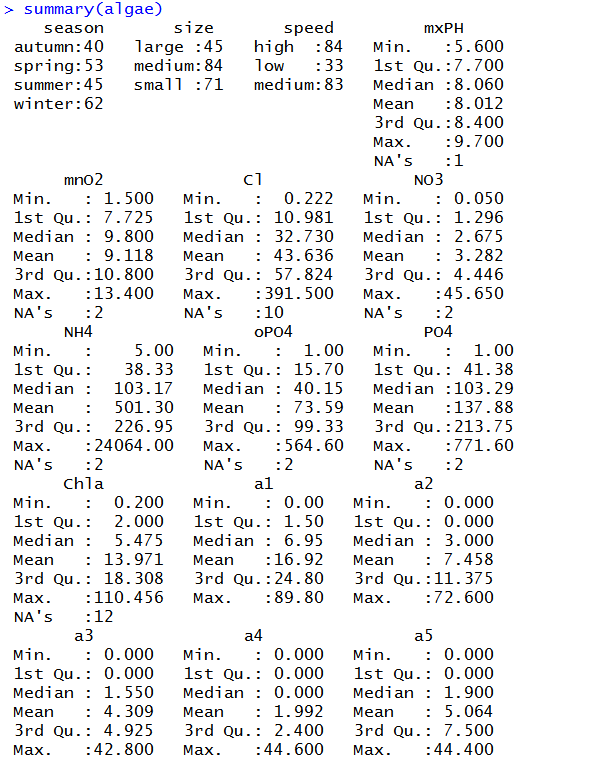
**1.Data Reading**

>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 algae data**



Summary of ‘algae’ Data

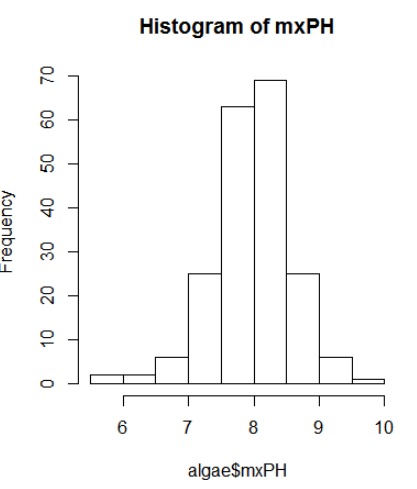


This simple function shows a statistical of the data summary for nominal variables, he gives the value of each variable frequency. For example, results from samples collected id winter than any other season in more rivers size to a maximum of 45, low river flow rates less. For a numeric variable, R provides 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, most of information can be pressed through graphics.

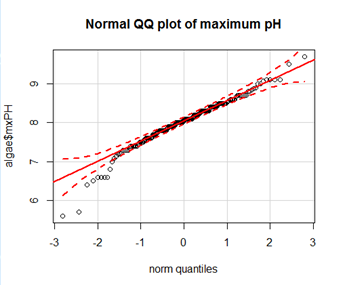
**2.Visualization**

(Draw a histogram)

>hist(algae$mxPH,main="Histogram of mxPH")



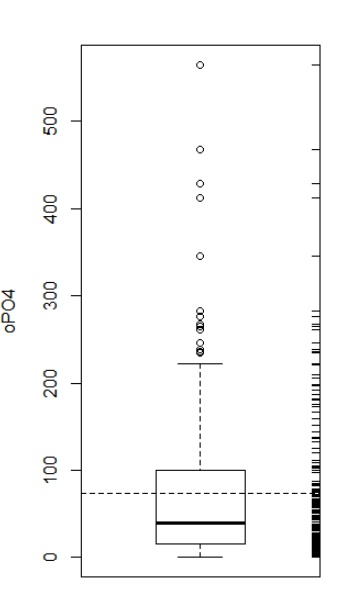
(Draw a QQ diagram)



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

Draw the variable values and theoretical quantiles of the normal distribution of scatter charts, 95%confidence interval for a given distribution of strip chart from above known variable has a few little value outside the 95% confidence interval, the y are nor subject to normal distribution

(Draw a box diagrams)



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

>rug(algae$oPO4,side=4)

>abline(h=mean(algae$oPO4),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.

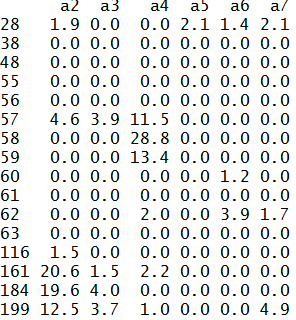
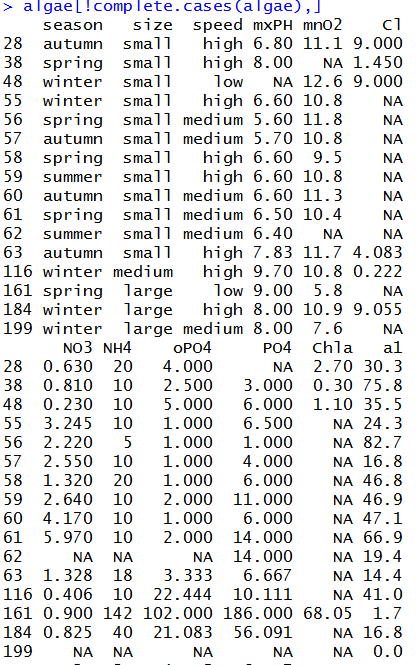
**3. Handling of missing data**

**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.

(Check the records with missing data)

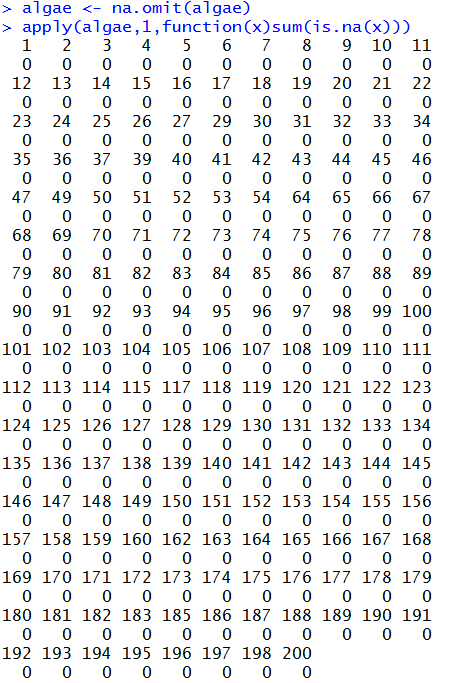
>algae[!complete.cases(algae),]



(Excluding missing values)

>algae <- na.omit(algae)

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



(Save the preprocessed data set)

>write.table(algae,file='Analysis1.txt',quote=FALSE,sep = '\t',row.names = FALSE,col.names = FALSE)

**3.2 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.

(Use centralImputation() to automatically fill data value with data center trends all missing values)

>library(DMwR)

>algae<-centralImputation(algae)

(Save the preprocessed data set)

>write.table(algae,file=’Analysis2.txt’,quote=FALSE,sep=’\t’,row.names=FALSE,col.names=FALSE)

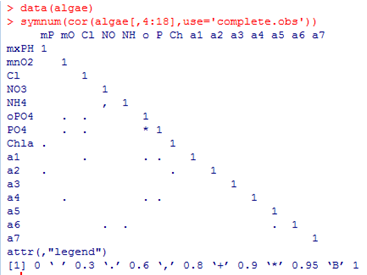
**3.3 By the correlation between property to fill in the missing values**

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.

(Gets the variables the correlation matrix)

>data(algae)

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



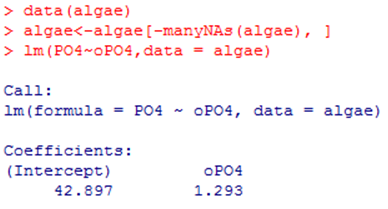
Results showed that most of the variables are not related, however, there are two exceptions: variable between NH4 and NO3, variable between PO4 and oPO4. Between the latter two variables related to a high value (greater than 0.9). And, therefore, identify missing data by them is dangerous. In addition, 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

(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 we variables are not at the same time has missing values, you can use this formular to calculate the missing values of these variables.

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

1.PNG

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

(Save the preprocessed data set)

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

**3.4Through the similarity between objects to fill missing data values**

Correlation between different from exploration data set 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. Second method is to use similar data weighted mean Weight size increases with distance from the outstanding cases of missing values for d, It weights in weighted average values for: w d = 𝑒−𝑑

1. Above method can add functions 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)

1. Save the preprocessed data set

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