****

**Beijing Institute Of Technology**

**School Name:** Computer Science & Technology

**Student Name:** HAQ IJAZ UL (2820150066)

**Course Name:** Data Mining

**Assignment**: Homework No1

**1: Load data**

Using R software to load DMWR add package, which has the name of the data frame we need algae. This data frame contains 200 observed values:

algae<-read.table("C:/Users/EJO/Desktop/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'))

head(algae)

season size speed mxPH mnO2 Cl NO3 NH4 oPO4

1 winter small medium 8.00 9.8 60.800 6.238 578.000 105.000

2 spring small medium 8.35 8.0 57.750 1.288 370.000 428.750

3 autumn small medium 8.10 11.4 40.020 5.330 346.667 125.667

4 spring small medium 8.07 4.8 77.364 2.302 98.182 61.182

5 autumn small medium 8.06 9.0 55.350 10.416 233.700 58.222

6 winter small high 8.25 13.1 65.750 9.248 430.000 18.250

PO4 Chla a1 a2 a3 a4 a5 a6 a7

1 170.000 50.0 0.0 0.0 0.0 0.0 34.2 8.3 0.0

2 558.750 1.3 1.4 7.6 4.8 1.9 6.7 0.0 2.1

3 187.057 15.6 3.3 53.6 1.9 0.0 0.0 0.0 9.7

4 138.700 1.4 3.1 41.0 18.9 0.0 1.4 0.0 1.4

5 97.580 10.5 9.2 2.9 7.5 0.0 7.5 4.1 1.0

6 56.667 28.4 15.1 14.6 1.4 0.0 22.5 12.6 2.9

Function head () will display the first 6 rows of the data frame. Each row of the data frame represents an observed value.

**2 Data summary**

Obtain descriptive statistics of the data.

> summary(algae)

season size speed mxPH

autumn:40 large :45 high :84 Min. :5.600

spring:53 medium:84 low :33 1st Qu.:7.700

season size speed mxPH

autumn:40 large :45 high :84 Min. :5.600

spring:53 medium:84 low :33 1st Qu.:7.700

summer:45 small :71 medium:83 Median :8.060

winter:62 Mean :8.012

3rd Qu.:8.400

Max. :9.700

NA's :1

mnO2 Cl NO3

Min. : 1.500 Min. : 0.222 Min. : 0.050

1st Qu.: 7.725 1st Qu.: 10.981 1st Qu.: 1.296

Median : 9.800 Median : 32.730 Median : 2.675

Mean : 9.118 Mean : 43.636 Mean : 3.282

3rd Qu.:10.800 3rd Qu.: 57.824 3rd Qu.: 4.446

Max. :13.400 Max. :391.500 Max. :45.650

NA's :2 NA's :10 NA's :2

NH4 oPO4 PO4

Min. : 5.00 Min. : 1.00 Min. : 1.00

1st Qu.: 38.33 1st Qu.: 15.70 1st Qu.: 41.38

Median : 103.17 Median : 40.15 Median :103.29

Mean : 501.30 Mean : 73.59 Mean :137.88

3rd Qu.: 226.95 3rd Qu.: 99.33 3rd Qu.:213.75

Max. :24064.00 Max. :564.60 Max. :771.60

NA's :2 NA's :2 NA's :2

Chla a1 a2

Min. : 0.200 Min. : 0.00 Min. : 0.000

1st Qu.: 2.000 1st Qu.: 1.50 1st Qu.: 0.000

Median : 5.475 Median : 6.95 Median : 3.000

Mean : 13.971 Mean :16.92 Mean : 7.458

3rd Qu.: 18.308 3rd Qu.:24.80 3rd Qu.:11.375

Max. :110.456 Max. :89.80 Max. :72.600

NA's :12

a3 a4 a5

Min. : 0.000 Min. : 0.000 Min. : 0.000

1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000

Median : 1.550 Median : 0.000 Median : 1.900

Mean : 4.309 Mean : 1.992 Mean : 5.064

3rd Qu.: 4.925 3rd Qu.: 2.400 3rd Qu.: 7.500

Max. :42.800 Max. :44.600 Max. :44.400

a6 a7

Min. : 0.000 Min. : 0.000

1st Qu.: 0.000 1st Qu.: 0.000

Median : 0.000 Median : 1.000

Mean : 5.964 Mean : 2.495

3rd Qu.: 6.925 3rd Qu.: 2.400

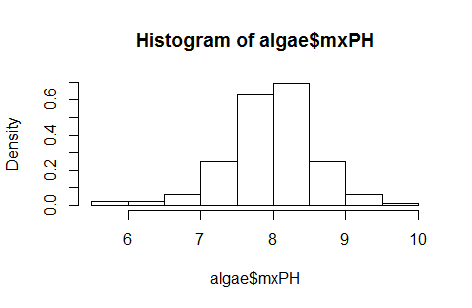
Max. :77.600 Max. :31.600

This simple function immediately gives the statistical characteristics of the data generalization, for the nominal variable, he gives the frequency of each value of the variable. For example, from the results we can know that the water samples collected in winter are more than other seasons, the river size for the maximum of 45, the river flow rate of low is relatively small. For the numeric variables, R provides us with a series of information about 1/4 digit, median, mean, 3/4 digit, extreme value and so on. These statistics provide the preliminary information about the distribution of values and in variables have missing values, Na strings behind the values for missing values of a number, the median, mean, quartile information, we can understand data distribution skewness and dispersion.

**3 Visualization of data**

> hist(algae$mxPH,probability = T)

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



> library(car)

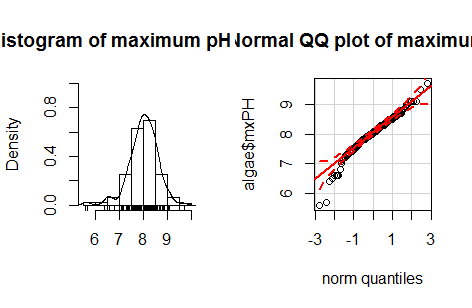
> par(mfrow=c(1,2))

> hist(algae$mxPH,prob=T,xlab = "",ylim = 0:1,main = "Histogram of maximum pH value")

> lines(density(algae$mxPH,na.rm=T))

> rug(jitter(algae$mxPH))

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



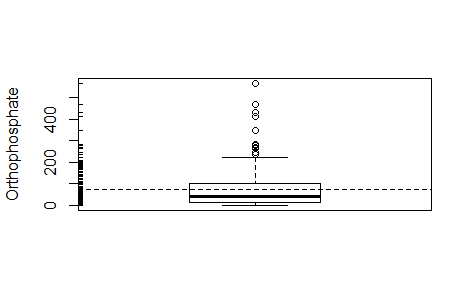
On the left side of the figure is the histogram of the kernel density curve, which can be easily distinguished from the actual value of the variable near the X axis, for example, we can find that there are two words that are significantly lower than all the other values. This kind of data is very important, 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.

Here is another example of the data test, using it to test the variable OP04:

> boxplot(algae$oPO4,ylab="Orthophosphate")

> rug(jitter(algae$oPO4),side = 2)

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



Represent values for variables of divergent and outliers and other special large value, usually considered to be outliers, which means that the box line graph is given a lot of information. It not only gives the variables of the central tendency are also given for the box top small horizontal line above the small circle.

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.

Sometimes, when there are outliers and need to determine which outlier observations, you can use the graphical method, such as drawing NH4 values, will pay attention to a particular value, we can use some methods to identify extreme values corresponding water samples:

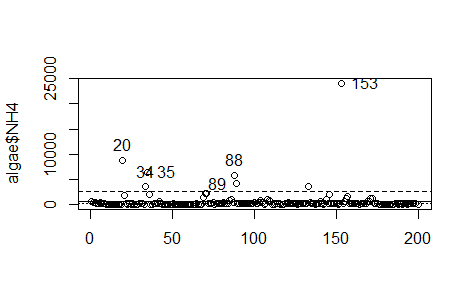
> plot(algae$NH4,xlab = "")

> abline(h=mean(algae$NH4,na.rm = T),lty=1)

> abline(h=mean(algae$NH4,na.rm = T)+sd(algae$NH4,na.rm = T),lty=2)

> abline(h=median(algae$NH4,na.rm = T),lty=3)

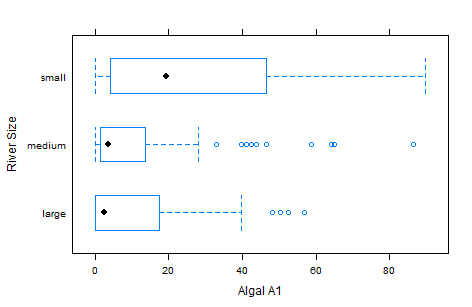
> identify(algae$NH4)



From the figure above know, sample of 153 NH4 value is a great outliers, 20, 34,35,88,89153 water samples from NH4 values are unusually high, can put the samples selected, look at them and other water algae have what kind of differences.

For different values of size, you can draw a set of variables A1 box chart.

bwplot(size~a1,data=algae,ylab = "River Size",xlab = "Algal A1")



From the map, in the smaller rivers, the higher the frequency of a1.

**4 Data deletion processing**

4.1 Deletion of missing parts

Eliminating records with missing data is very easy to achieve, especially when the proportion of these records in the available data set is very small, the choice is more reasonable.

It is best to check the observed value, or at least the number of the observed value, if at least one of the missing data is removed, for example:

algae[!complete.cases(algae),]

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

[1] 16 missing samples

Function complete.cases () produces a Boolean vector, the vector elements and the number of algae data frame in the same number of, if the corresponding row in the data frame does not contain Na values (i.e., for a complete observation value), the function returns are the true values. The aforementioned "!" operator, it is logical not, so the above instructions show that contains missing records of water samples.

In order to remove the 16 samples from the data frame, we can simply input:

algae <- na.omit(algae)

Even if we decide not to use extreme methods that eliminate all missing values, we can eliminate some of the observed values. Because the missing value of these samples is too much, so they are almost useless samples, if the use of sophisticated methods to fill the missing values, it will lead to a larger deviation. Need to pay attention to is that if the implementation of the previous command, you need to re read the data, because the instruction has removed all of the missing data, so the next command is meaningless! Observation in the sample data, we can see article 62 and article 199 record of 11 variables, six missing values. In this case, it's better to remove them:

Algae left algae[-c (62,199)].

In some problems, it is not feasible to check the missing values of the data with the above observation method because of missing values in a large number of records, so it is necessary to find out the lines with more missing values. The following code can find the missing values of the algae data set for each line of data

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

Function apply () is a very powerful function of R in a class of functions. This kind of function is also called the meta function, which can be applied to other functions in some conditions. For a function apply, it applies a function to each row of the data frame. Given in the third parameter to apply () function for this application is function, for each row of the data frame respectively call the function. In this case, we use a temporary function. It only exists when the apply () function is called. In addition, the function apply () of the third parameters can be a normal function of the name of the function. The function of a temporary function is to calculate the number of NA in the object. In R, the logical value TRUE is equal to the value 1, the logical value FALSE is equal to the value 0, which means that when adding a Boolean value vector, the number of elements in the vector is obtained.

According to the above code, you can write a program to find the algae contains a given number of missing values. The function can be used as follows:

data(algae)

manyNAs(algae,0.2)

[1] 62 199

Only a few lines of data with missing values are removed from the previous operation will be required to call the function data (). Function manyNA is () the function is to find the missing value is greater than the number of rows in the column 20%. A precise number of columns can be set in the second parameter. Therefore, use the following code to do not need to know the specific number of rows with missing values.

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

The above operation can remove the sample data of 62199 water samples.

4.2 To fill the missing value with a high frequency value.

sample algae[48,] the variable mxPH has missing values. Because the distribution of this variable is approximately normal distribution, we can choose the average value to fill the hole, and the calculation method is as follows:

algae[48,'mxPH'] <- mean(algae$mxPH,na.rm=T)

Here, the function mean () to calculate the average value of the vector, the parameters of the na.rm=T to ignore the missing data.

Most often used to fill a column in all the missing values rather than one line and one line to fill. Taking the variable Chla as an example, this variable has a missing value on the twelfth row. In addition, this is also an average value can not represent the value of most of the variables. In fact, the distribution of Chla tends to lower values, and it has several extreme values, which make the mean value cannot represent most of the variable values. Therefore, we use the median to fill the missing values of this class:

algae[is.na(algae$Chla),'Chla'] <- median(algae$Chla,na.rm=T)

Because the missing value can cause some methods to be used, it is generally considered not a good method to use the above method to fill in the missing values. Although the above methods are fast, especially suitable for large data sets, but it may lead to large data deviation, the impact of the late data analysis. However, finding the optimal data with no error method is complex and may not be applicable to large data mining problems.

4.3 to adjust the missing value through the correlation of variables

Another way to obtain the missing value of the missing value is to explore the correlation between variables. For example, through the correlation between the values of the variables, can be found in a high degree of correlation between the variables and mxPH. This allows us to get the more likely to fill the missing values of the forty-eighth samples. This is before the use of the average method will be better.

> options(digits = 1)

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

mxPH mnO2 Cl NO3 NH4 oPO4 PO4 Chla a1

mxPH 1.00 -0.10 0.15 -0.17 -0.15 0.090 0.10 0.43 -0.16

mnO2 -0.10 1.00 -0.26 0.12 -0.08 -0.394 -0.46 -0.13 0.25

Cl 0.15 -0.26 1.00 0.21 0.07 0.379 0.45 0.14 -0.36

NO3 -0.17 0.12 0.21 1.00 0.72 0.133 0.16 0.15 -0.25

NH4 -0.15 -0.08 0.07 0.72 1.00 0.219 0.20 0.09 -0.12

oPO4 0.09 -0.39 0.38 0.13 0.22 1.000 0.91 0.11 -0.39

PO4 0.10 -0.46 0.45 0.16 0.20 0.912 1.00 0.25 -0.46

Chla 0.43 -0.13 0.14 0.15 0.09 0.107 0.25 1.00 -0.27

a1 -0.16 0.25 -0.36 -0.25 -0.12 -0.395 -0.46 -0.27 1.00

a2 0.34 -0.07 0.08 0.02 -0.04 0.124 0.13 0.37 -0.26

a3 -0.03 -0.24 0.08 -0.09 -0.11 0.006 0.03 -0.06 -0.11

a4 -0.18 -0.38 0.14 -0.01 0.27 0.382 0.41 -0.09 -0.09

a5 -0.11 0.21 0.15 0.21 0.02 0.122 0.16 -0.07 -0.27

a6 -0.17 0.19 0.17 0.54 0.40 0.003 0.05 0.01 -0.26

a7 -0.17 -0.10 -0.04 0.08 -0.03 0.026 0.08 0.02 -0.19

a2 a3 a4 a5 a6 a7

mxPH 0.34 -0.027 -0.18 -0.11 -0.173 -0.17

mnO2 -0.07 -0.235 -0.38 0.21 0.189 -0.10

Cl 0.08 0.077 0.14 0.15 0.169 -0.04

NO3 0.02 -0.092 -0.01 0.21 0.544 0.08

NH4 -0.04 -0.113 0.27 0.02 0.401 -0.03

oPO4 0.12 0.006 0.38 0.12 0.003 0.03

PO4 0.13 0.032 0.41 0.16 0.053 0.08

Chla 0.37 -0.063 -0.09 -0.07 0.010 0.02

a1 -0.26 -0.108 -0.09 -0.27 -0.262 -0.19

a2 1.00 0.010 -0.18 -0.19 -0.134 0.04

a3 0.01 1.000 0.03 -0.14 -0.197 0.04

a4 -0.18 0.033 1.00 -0.10 -0.085 0.07

a5 -0.19 -0.142 -0.10 1.00 0.389 -0.05

a6 -0.13 -0.197 -0.08 0.39 1.000 -0.03

a7 0.04 0.039 0.07 -0.05 -0.030 1.00

The function cor () function is to generate the correlation between the variables of the matrix (because the first 3 variables when the nominal variable, so do not consider them when calculating the correlation value). When setting the parameter use= "complete.obs", the R ignores the record containing the NA when calculating the correlation value. Correlation values between 1 (or -1) indicated that there was a strong positive (or negative) linear correlation between the two variables. Then the other R functions can be obtained in the form of linear correlation between variables, which allows us to calculate the value of the other variable through the value of a variable.

Function cor () the output result is not very clear, can be used to improve the results of the output form of symnum.

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

mP mO Cl NO NH o P Ch a1 a2 a3 a4 a5 a6 a7

mxPH 1

mnO2 1

Cl 1

NO3 1

NH4 , 1

oPO4 . . 1

PO4 . . \* 1

Chla . 1

a1 . . . 1

a2 . . 1

a3 1

a4 . . . 1

a5 1

a6 . . . 1

a7 1

attr(,"legend")

[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘\*’ 0.9 ‘B’ 1

In the data of this case, most of the variables are not related, however, there are two exceptions: the variables NH4 and NO3, between the variables PO4 and oPO4. After the correlation between the two variables is very high (greater than 0.9). 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. In order to achieve this goal, we need to find a linear relationship between the two variables, and the method is as follows:

> data(algae)

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

> lm(PO4~oPO4,data = algae)

Call:

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

Coefficients:

(Intercept) oPO4

42.90 1.29

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

After removing the sample 62 and sample 199, there is still a sample 28 missing value on the PO4, you can simply use the above linear relationship to calculate the missing value of the fill value:

> algae[28,'PO4'] <- 42.897 + 1.293 \* algae[28,'oPO4']

> algae[28,"PO4"]

[1] 48

Through the correlation between the variables to find out the water sample 28, PO4 missing value to fill 48.

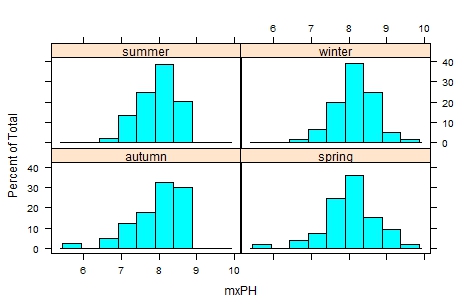
Research on linear relationships allows us to fill some of the missing values. However,

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

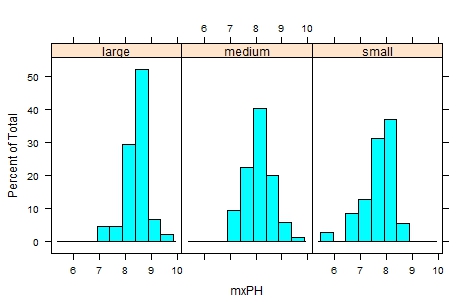
histogram. Drawing code is as follows:

> 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)



Then 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:

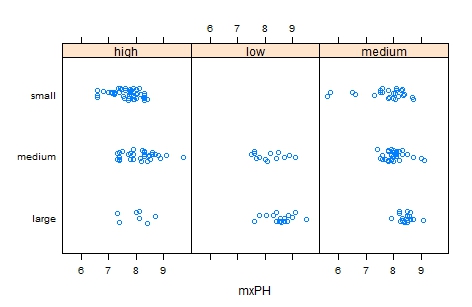
> stripplot(size~mxPH|speed,data=algae,jitter=T)

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.

****

4.4 to fill the missing values by exploring the similarities between the cases

The correlation between the data set is different from the exploration, try to use the line between similarity to fill in the missing value. We can use this method to fill out the two missing data that contains too many NA that are worth out of sample.

As described above, if the two water samples are similar, some of the variables have missing values, then the missing value is likely to be similar to the value of another water sample. In order to use this intuitive method, first define the concept of similarity. Similarity is often defined by variables that describe the multivariate metric space of the observed value. In the literature there are many metrics of similarity, which are commonly used in Euclidean distance. This distance can be informally defined as the sum of squares of the difference between any two case observations, and the formula is as follows:

The following description of the method is to use this measure to find the most similar to any of the 10 water samples with missing values, and to fill the missing values with them. We consider two methods for applying these values. The first method simply calculates the median of the 10 most similar cases and fills them with the median. The second method is to use the weighted mean value of these similar data. The weight of the size of the distance to be filled with the missing value of the distance to the case of D, then its value in the weighted average weight is:

w (d) =e^ (-d)

The method above can be achieved by adding a packet function knnImputation (), a variant of this function with a Euclidean distance to find the nearest k neighbor of any case. When calculating the distance, the data should be standardized.

The following describes how to use the knnImputation () function

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

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

> algae

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

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

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

algae[!complete.cases(algae),]

