数据挖掘作业一

姓名：邢奕鹏 学号：2520150263

数据摘要

1.对标称属性，给出每个可能取值的频数，

2. 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。

1.

freSeason = tabulate(D{1});

freSize = tabulate(D{2});

freSpeed = tabulate(D{3});

'winter' [62] [ 31]

'spring' [53] [26.5000]

'autumn' [40] [ 20]

'summer' [45] [22.5000]

'small' [71] [35.5000]

'medium' [84] [ 42]

'large' [45] [22.5000]

'medium' [83] [41.5000]

'high' [84] [ 42]

'low' [33] [16.5000]

2.

Value(:,i) = str2double(value(:,i));

Max(i,1) = max(Value(:,i));

Min(i,1) = min(Value(:,i));

Mean(i,1) = mean(Value(:,i));

Median(i,1) = median(Value(:,1));

Q1(i,1) = prctile(Value(:,1),25);

Q3 (i,1)= prctile(Value(:,1),75);

9.70000000000000

13.4000000000000

391.500000000000

45.6500000000000

24064

564.599980000000

771.599980000000

110.456000000000

89.8000000000000

72.6000000000000

42.8000000000000

44.6000000000000

44.4000000000000

77.6000000000000

31.6000000000000

上面一列是不同数值属性的最大值，最小值、均值、中位数、四分位数均可以同样现实出来。

数据的可视化

针对数值属性，

* 绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。
* 绘制盒图，对离群值进行识别

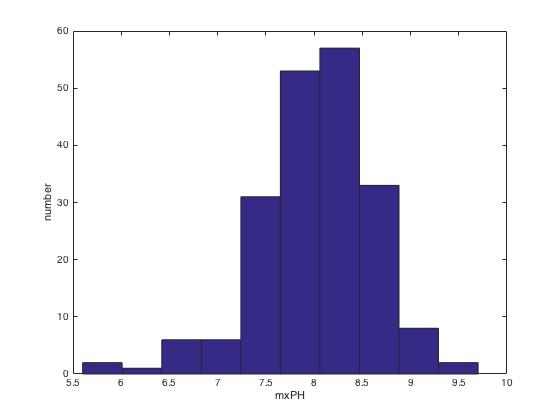
对7种海藻，分别绘制其数量与标称变量，如size的条件盒图

1.

hist(Value(:,1));

xlabel('mxPH')

ylabel('number')



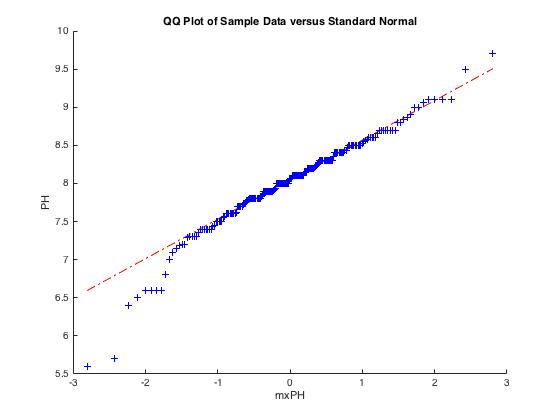
图一

图一是mxPH的直方图。

qqplot(Value(:,1));

xlabel('mxPH')

ylabel('PH')



图二

图二是qq图检验是否为正态分布

2.

boxplot(Value(:,1));

xlabel('mxPH')

ylabel('PH')

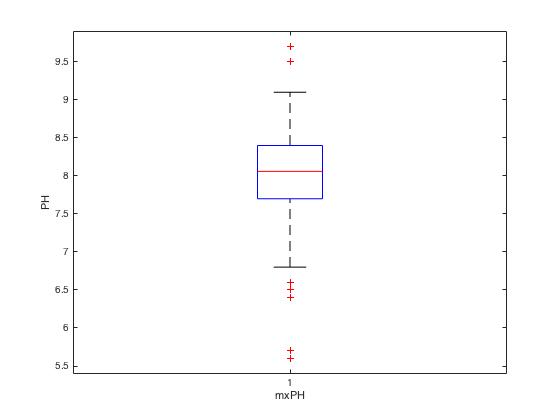


图3

图3是绘制的mxPH的盒图，图中可以看出离群值。