实验报告

1 实验内容

有 200 个水样，每条记录是同一条河流在该年的同一个季节的三个月内收集的水样的平均值。每条记录由 11 个变量构成，3 个是标称变量，分别描述水样收集的季节，河流大小和河水速度，剩下的 8 个变量是水样的化学参数：

•最大 pH 值(mxPH)
•最小含氧量(mnO2)
•平均氯化物含量(Cl)
•平均硝酸盐含量(NO3)
•平均氨含量(NH4)
•平均正磷酸盐含量(oPO4)
•平均磷酸盐含量(PO4)
•平均叶绿素含量(Chla)

a1-a7 为 7 种不同有害藻类在相应水样中的频率数目。

要求

对标称属性，给出每个可能取值的频数，
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。
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

数据缺失的处理

分别使用下列四种策略对缺失值进行处理:

•将缺失部分剔除
•用最高频率值来填补缺失值
•通过属性的相关关系来填补缺失值
•通过数据对象之间的相似性来填补缺失值

实验过程

首先是获取标称属性的频率

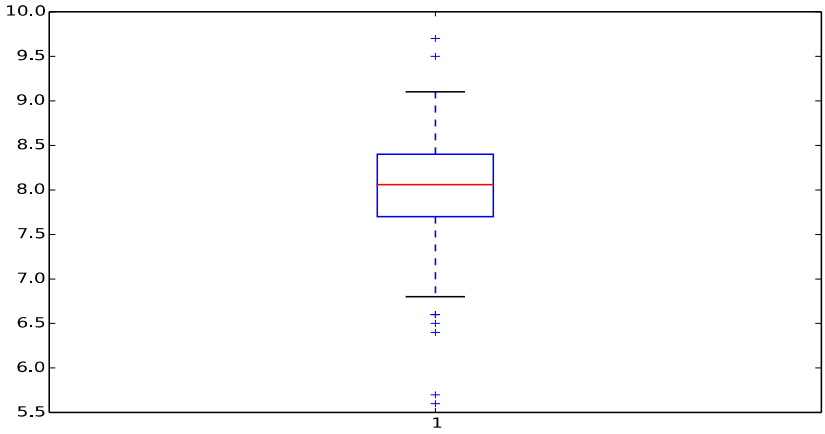
spring 53
summer 45
autumn 40
winter 62
small 71
medium 84
large 45
low 33
medium 83
high 84

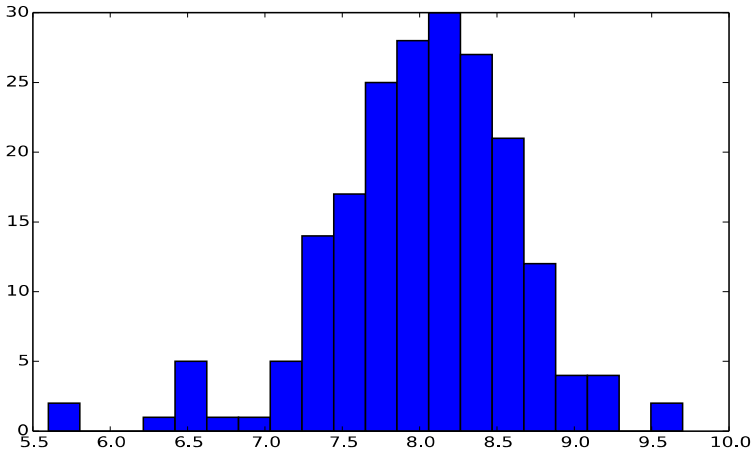
接下来，数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数

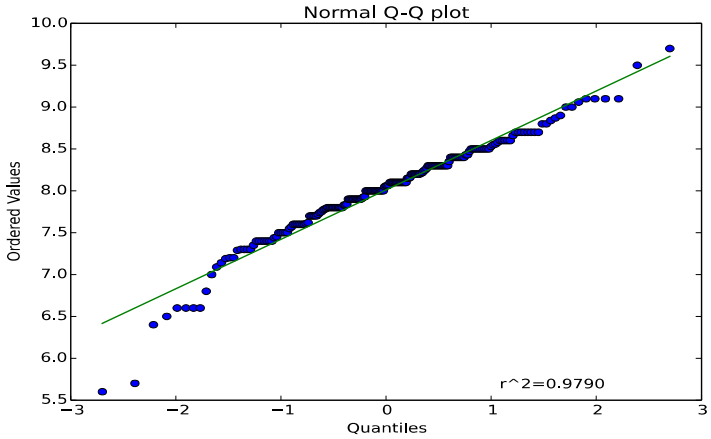
mxPH 9.7 5.6 8.01173366834 8.05 7.7 8.4 1
mnO2 13.4 1.5 9.11777777778 9.8 7.725 10.8 2
Cl 391.5 0.222 43.6362788421 32.73 10.98125 57.8235 10
NO3 45.65 0.05 3.28238888889 2.675 1.296 4.44625 2
NH4 24064.0 5.0 501.295828384 103.1665 38.33325 226.9500075 2
oPO4 564.59998 1.0 73.5905959596 40.15 15.7 99.33325 2
PO4 771.59998 1.0 137.88210096 103.2855 41.37525 213.75 2
Chla 110.456 0.2 13.9711968085 5.475 2.0 18.3075 12
a1 89.8 0.0 16.9235 6.95 1.5 24.8 0
a2 72.6 0.0 7.4585 3.0 0.0 11.375 0
a3 42.8 0.0 4.3095 1.55 0.0 4.925 0
a4 44.6 0.0 1.9925 0.0 0.0 2.4 0
a5 44.4 0.0 5.0645 1.9 0.0 7.5 0
a6 77.6 0.0 5.964 0.0 0.0 6.925 0
a7 31.6 0.0 2.4955 1.0 0.0 2.4 0

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

拿其中的最大 PH 值来说，盒图，直方图，以及 QQ 图如下，通过 QQ 明显可以看到，该数据服从正态分布。

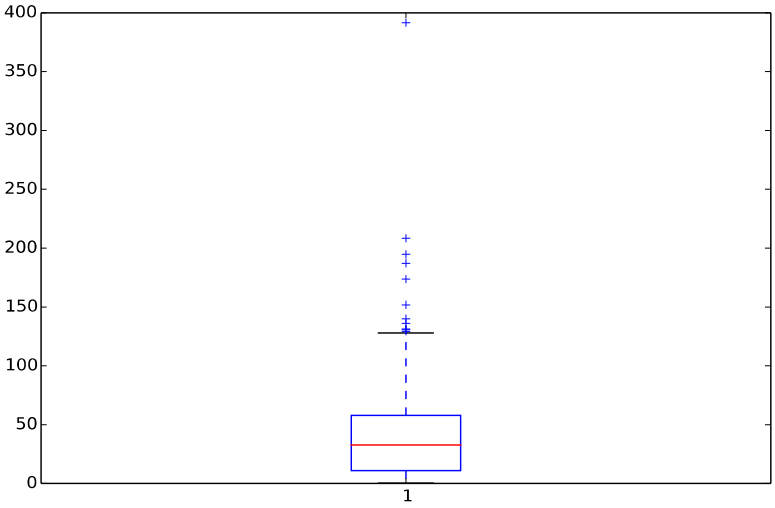

10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
1

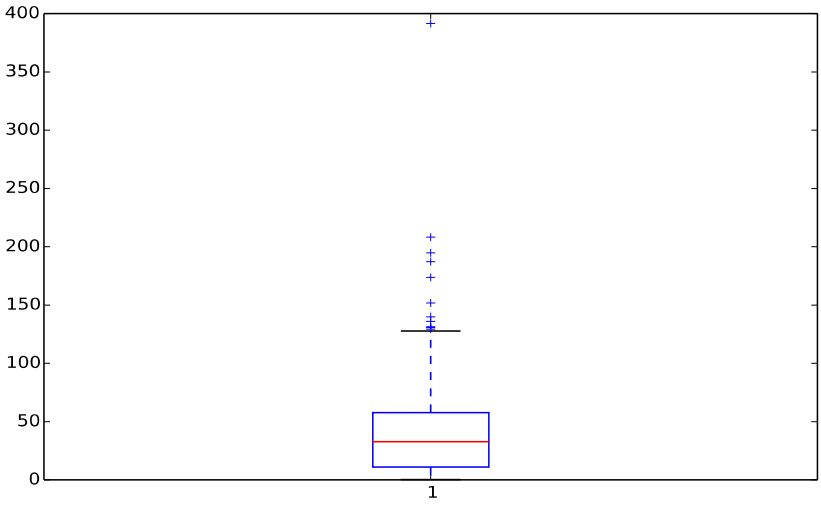

30
25
20
15
10
5
0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0


Normal Q-Q plot
Ordered Values
Quantiles
r^2=0.9790
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
−3
−2
−1
0
1
2
3

对于缺失值的处理，最开始实现的就是将却是的数据丢掉，即只要该数据为“XXXXXX”，即删掉该条记录。

接下来使用的是最高频率的数代替缺失值，使用数据可视化之后，发现两者效果并不是太明显。例如，Cl 属性缺失的值最多，但是采用该方法之后，看起来也基本上没有太大差别。





通过属性的相关关系来填补缺失值，可以计算各个属性之间的协方差，取其中相关性最大的属性，目前只实现了求协方差，还没有实现怎么填补。

通过属性的相似性来填补缺失值。基本思路是计算各个数据之间的欧式距离，距离最小则最相似。利用相似的数据属性值来填补缺失的值。在计算相似性之前，先将数据进行归一化，采用（value-min）/（max-min）的方式，标称变量不参与运算。