海藻数据分析实验报告

姓名：唐 育 洋

学号：2120151034

# 实验目的

数据预处理和分析是数据挖掘中的一个重要环节，通过本次实验了解和认识数据挖掘中的数据预处理和数据分析基本知识和掌握相关技能。

# 实验内容

实验以一份海藻数据为背景，主要完成一下两个实验内容：

(1)、获取数据摘要和可视化

(2)、对缺失属性值进行四种策略的处理，并对比。

数据集的信息如下：该数据集包含200条数据，每条数据共18个维度，前三个为非数值性，分别为season、size、speed。后15个为数值属性，其中第4到第11个属性代表化学属性，意义如下：

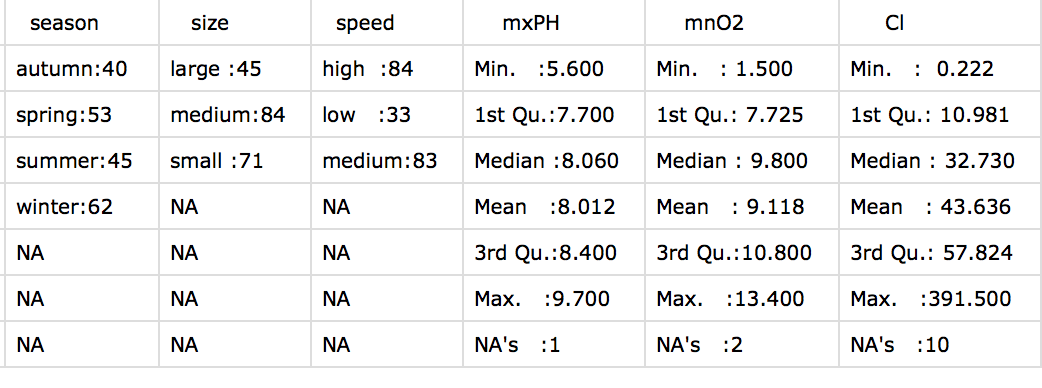
* 最大pH值(mxPH)
* 最小含氧量(mnO2)
* 平均氯化物含量(Cl)
* 平均硝酸盐含量(NO3)
* 平均氨含量(NH4)
* 平均正磷酸盐含量(oPO4)
* 平均磷酸盐含量(PO4)
* 平均叶绿素含量(Chla)

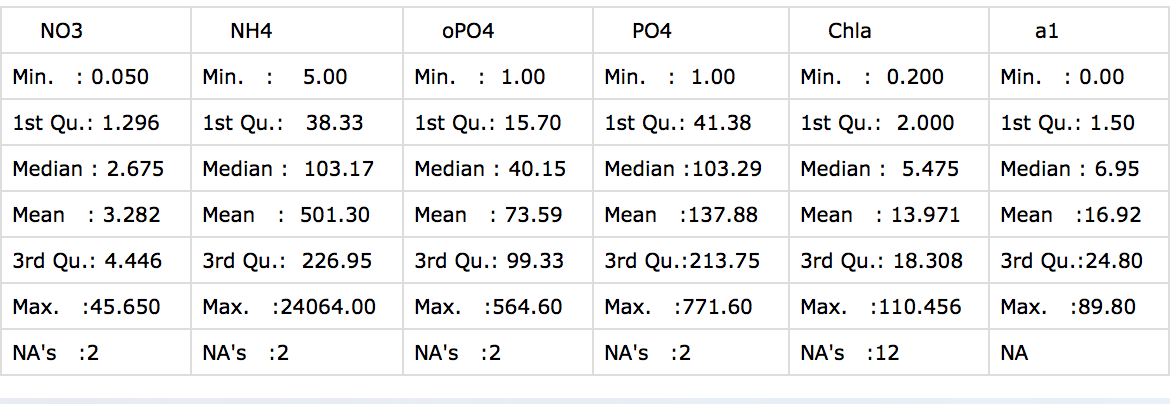
最后8个为藻类出现情况，用a1-a7表示。

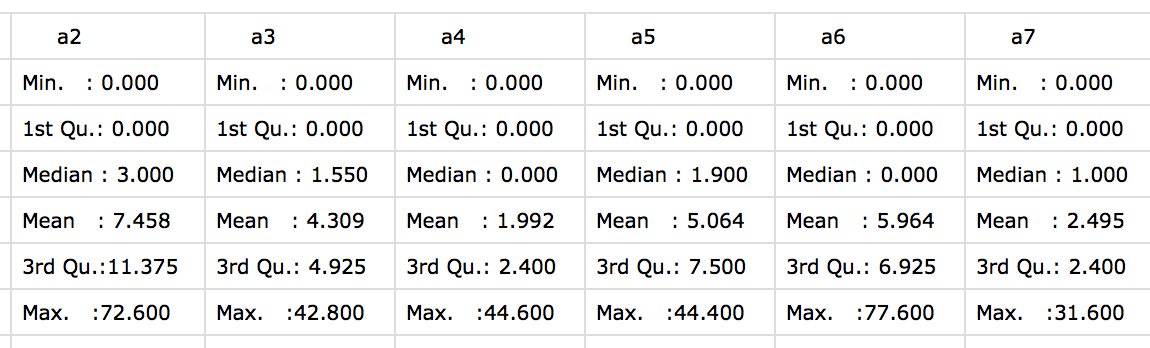
# 实验步骤

1、对于原始数据进行概要描述。

在接触到一个数据处理任务之前，我们一般对这个领域缺少先验知识，因此，对数据有个概要了解是非常有必要的。概要信息包括数值属性的最大值、最小值、中值、均值、众数、四分位数。对于非数值属性，一般指各值的出现频率和最高频率值。另外，对于属性的缺失值需要进行统计，这对我们分析数据剔除噪声数据、补全属性值等具有重要意义。本次实验采用的是R语言进行数据预处理和分析，使用summary函数可以得到，如下图所示：





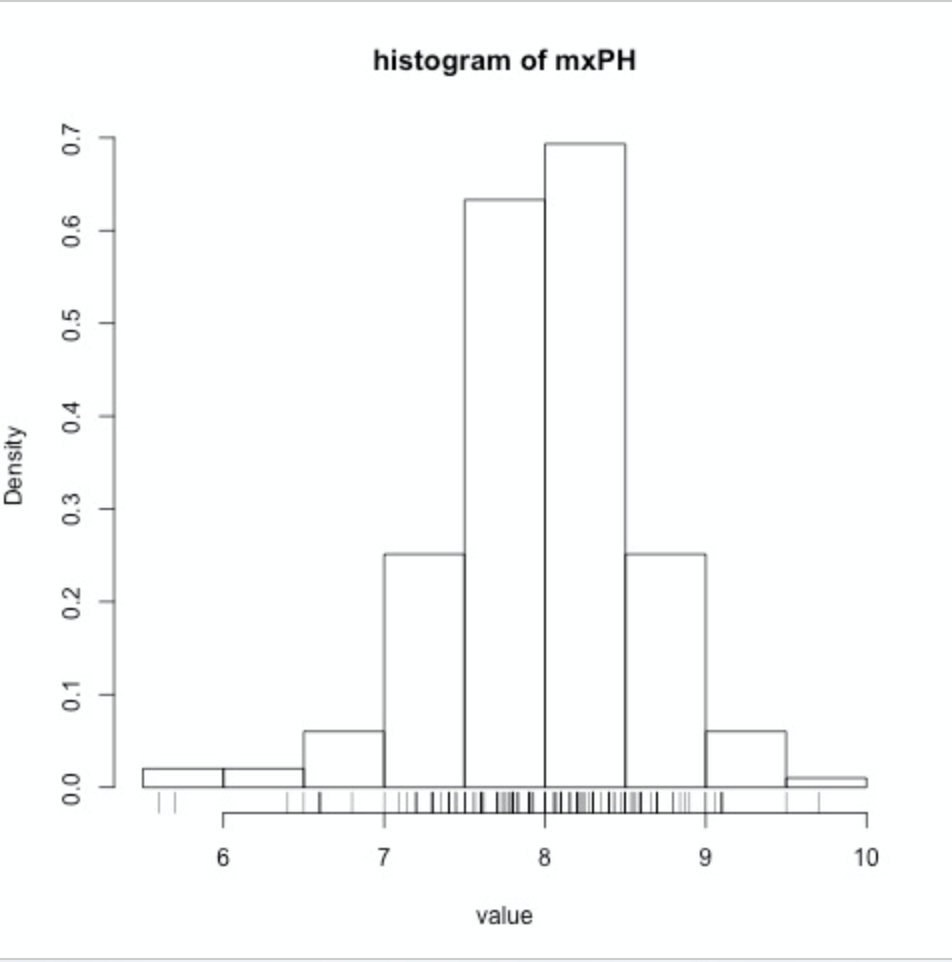


图中，season、size、speed是非数值属性，因此列出了各属性的取值及频率，剩下的为属性值，如前所述，给出了六项统计值。我们可以看出，来自冬季的测量数据比较多，有6个数据缺失Cl值，12个数据缺失Chla值，等信息，为我们初步了解数据提供有用信息。

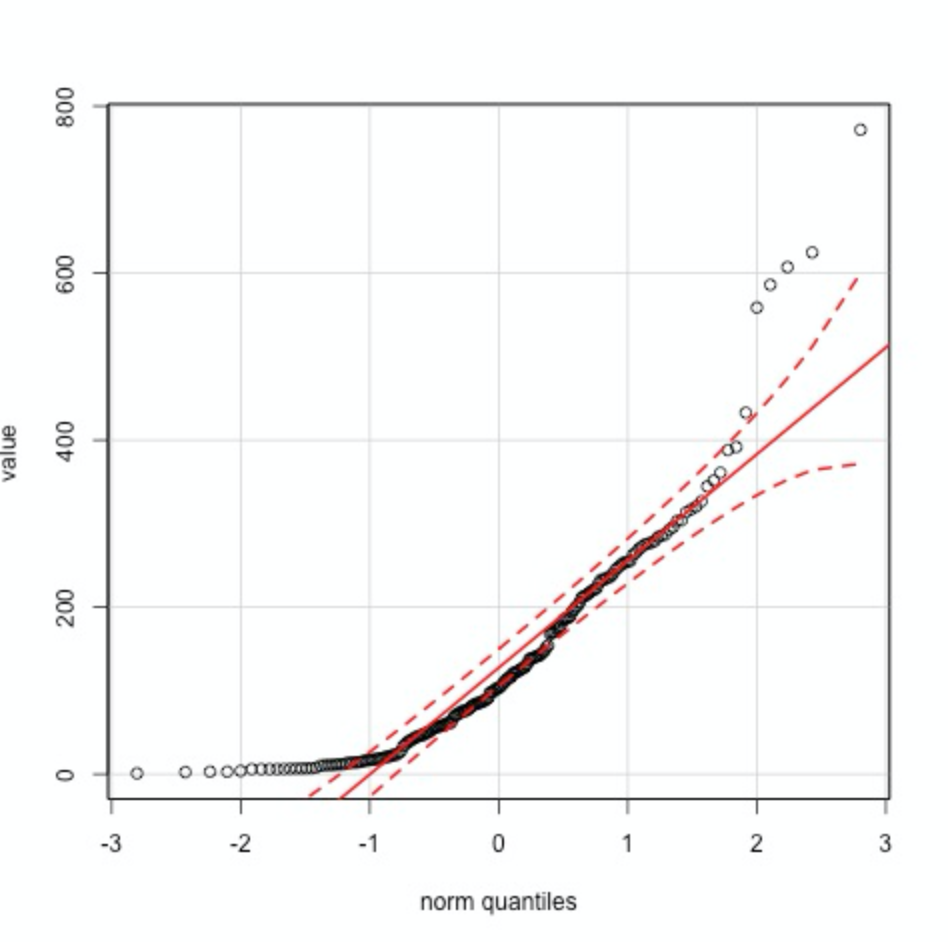
2、对各个属性进行可视化

我们将利用直方图、盒图等对数据进一步了解和认识。

对于数值属性，我们可以画出该属性的数据直方图，例如：

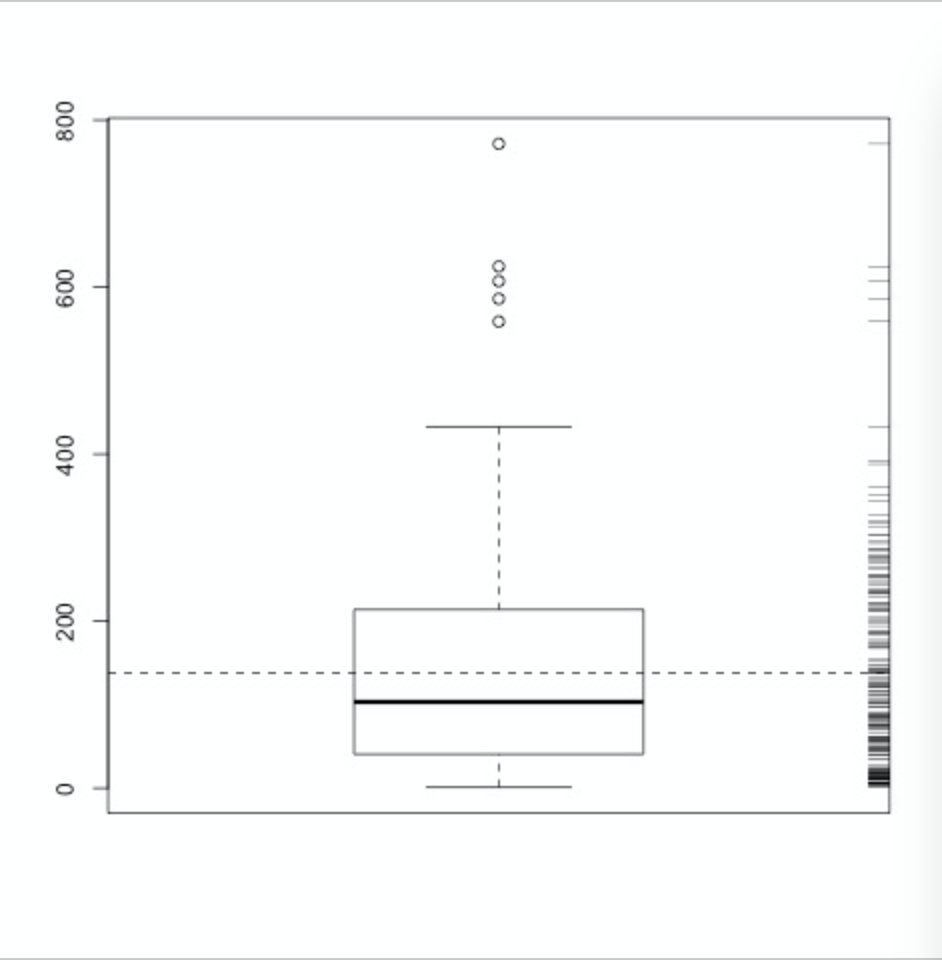


通过直方图，我们可以看到大部分大部分数据在该属性上的分布范围。同是，我们可以做一个QQ图，用于作数据和正态分布理论分位上的比较，如下：



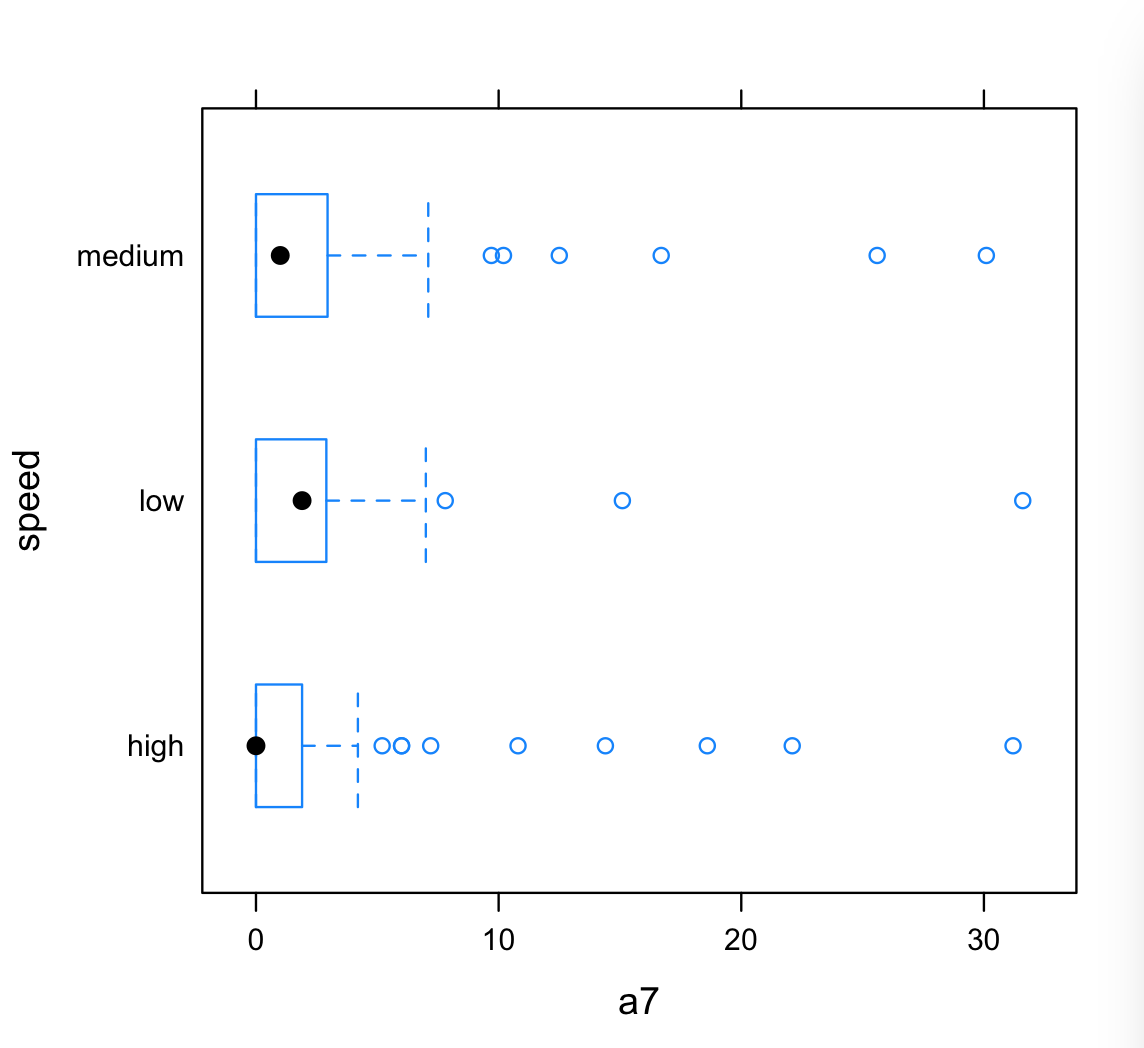
我们可以看到，图中红色虚线给出了95%置信区间的范围，在范围之外的，属于此分布的小概率事件，即这些数据不符合正态分布。

然后，我们画了各属性的盒图。通过盒图，我们可以更清晰直观地看到属性的分布，以及离散点的分布。如下图所示：



图中，水平盒子的上下边界代表数据的两个四分位点Q1、Q3，Q1与Q3的差值为IQR，即IQR=Q3-Q1，盒子之外的两条水平直线代表观测值的最小值和最大值，取值分别是Q1-1.5\*IQR和Q3+1.5\*IQR，粗体的水平线为中位数，圆圈代表着离散点。此外，我们也添加了一条水平虚线，代表着均值，数据的实际分布在右侧标上，作为对应。通过此图可以更加直观地发现离散点的分布状况。

此外，可以通过条件盒图可以变现某个名义变量如何影响两一个变量。如下图



3、缺失数据处理

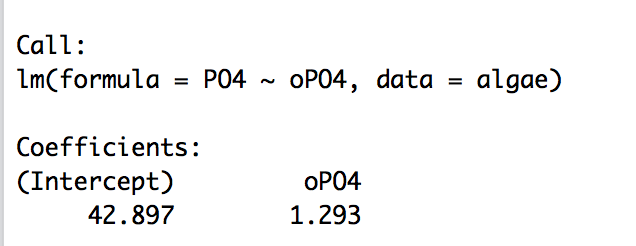
数据集中难免有些数据的属性值有缺失，这在实际应用中是很常见的，本次实验中，针对缺失数据，采用四种策略分别进行处理，并最后进行对比。四种策略分别是：

* 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

第一种方法最为直接简单，直接去掉有缺失属性的数据，这种方法适合具有少量缺失的情况，直接去掉之前一般需要对数据进行一次检查。此方法的缺点是，一旦缺失属性的数据多了，直接去掉将对后面的数据分析和数据挖掘产生不好作用。

第二种方法顾名思义，通过最高频率值来替换缺失值，此外也可以通过使用中位值、均值等来进行替换。

第三种方法利用属性变量之间的相关性进行补全。首先去除缺省情况太多的数据，然后求出所有数值属性之间的相关性矩阵，本次实验中，将此矩阵可视化可以较为明显地看到PO4与oPO4具有较强相关性。再在这两个模型之间建立回归模型，例如这里两个变量，选择的是线性回归模型，得到最优模型参数，通过该模型值为缺失的这两个属性进行补全。如下图所示，我们可以得到两者之间的关系为：PO4 = 1.29\*oPO4 + 42.90。



第四种方法利用数据对象之间的相似性进行补全，在数据集中寻找最为相似的数据，按照它的值来补全缺失值。这个思路扩展一下，也可以采用寻找最为接近的k个数据，按照一定权值进行组合。

得到这四个处理后的数据集后，保存了处理后的数据集，做了可视化，我们可以与原数据集进行对比。

# 实验收获

通过本次实验，学到了数据预处理和分析的基本方法和步骤，学会了如何使用R语言进行数据预处理和分析。