马的疝病分析

姓名：李丝雨

学号：2120161010

## 问题描述

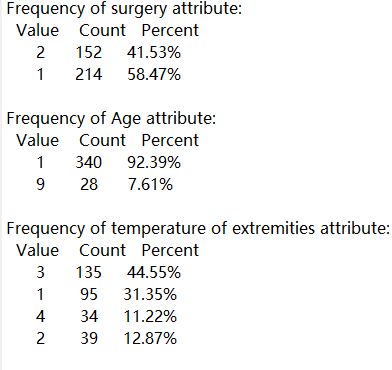
疝病是描述马胃肠痛的术语，这种病不一定源自马的胃肠问题，其他问题也可能引发马疝病。所给数据集是医院检测的一些指标。

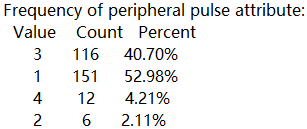
## 数据分析

分析程序的具体说明请参考readme。程序的编写语言是Matlab，编写平台是Matlab R2016b，共包含5个文件。

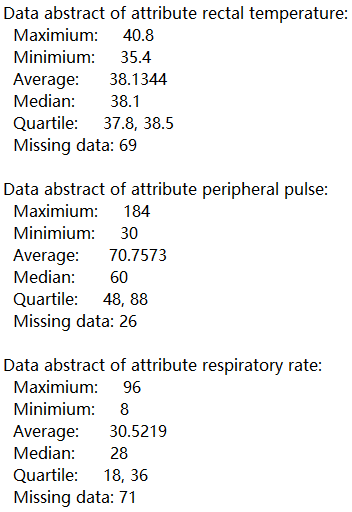
### 2.1数据摘要

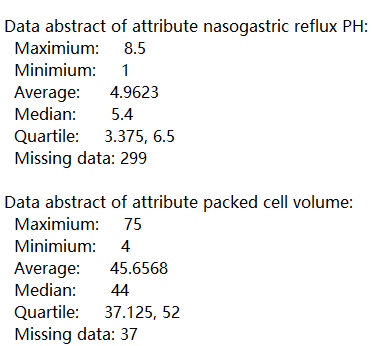
利用分析程序对数据进行了摘要以及可视化处理，数据摘要结果如下：首先为标称属性的频数统计结果，共统计了20个标称属性（hospital number属性没有统计，因为每个样本都不同，没有统计的必要），统计结果如下（由于数量太多，未一一列出，举例显示）：

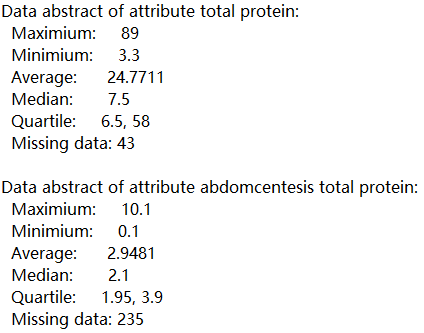




对于数值属性，程序统计了其最大、最小、均值、中位数、两个四分位数以及缺失数据的个数。结果如下列运行结果截图所示：

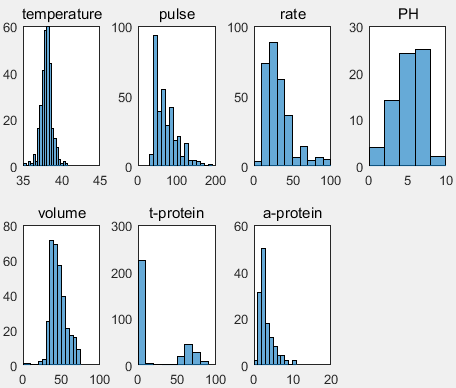






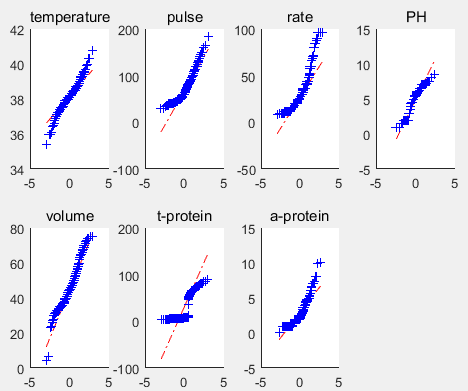
### 2.2数据可视化

对数值属性数据进行可视化处理，首先显示各个数值属性的直方图：



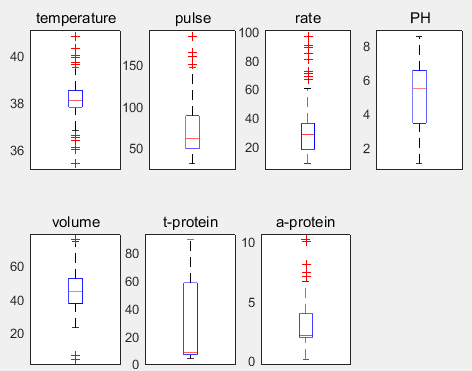
能看出rectal temperature 和nasogastric reflux PH属性相对近似于正态分布，其余属性值大部分集中在值较低的区间，total protein属性具有很好的可区分性。

再观察属性的QQ图：



能够看出前两个属性以及第五个属性拟合正态分布较好。

7个属性的盒图为：



可以看出rectal temperature属性和respiratory rate属性相比较于其他属性的盒图具有较多的离群值。

## 数据集预处理

对数据进行4种处理，分别为剔除缺省数据、按最高频率填补值、按属性填补以及按样本相似性填补。四种方式的做法为：

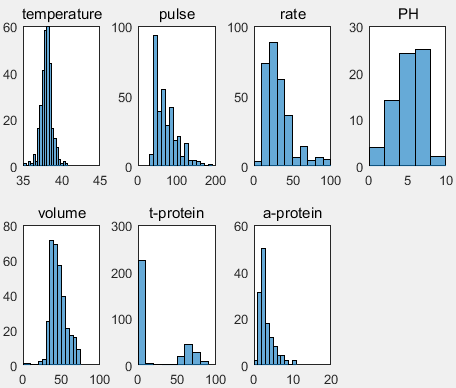
剔除缺省数据：将每个样本的属性进行核查，删除有缺省属性的样本；

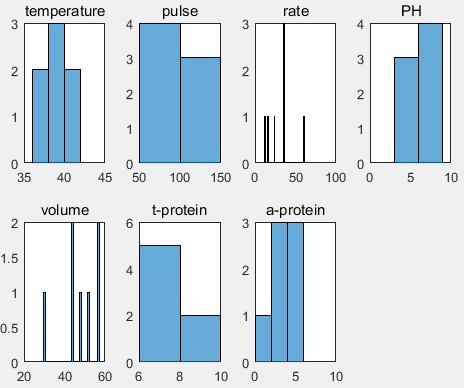
按最高频率值：对于某一属性下的缺失，使用该属性的最高频率值代替缺失的数据；

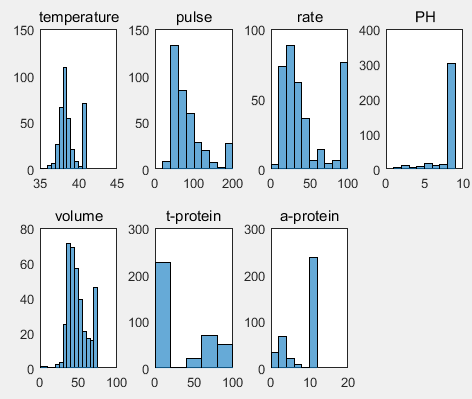
按属性填补：计算两个属性的相关性，相关性越大表明可以根据另一属性推断缺失属性的值。通过另一属性的回归分析，计算当前的缺失值；

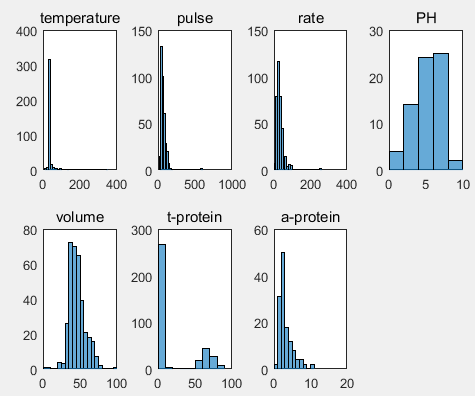
按相似性填补：计算两个样本的相似程度，越相似证明越可以使用该样本推断当前含缺失值的样本。

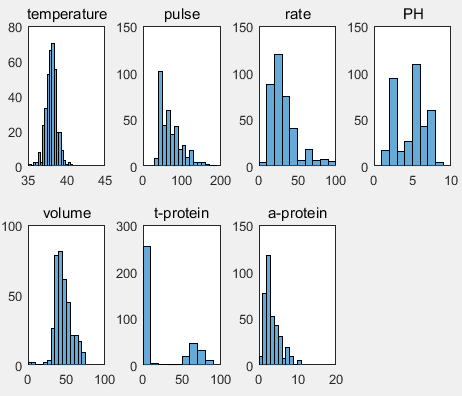
预处理后的数据集也可以进行可视化，以直方图为例，如下列图所示，其中第一个为处理前的直方图，后面依次为剔除缺省数据，按最高频率值填补，按属性填补，按相似性填补后的结果:











可以看到，剔除法剩余的样本数量太少，直方图所包含的信息量不足；按相似性填补的结果和处理前更相似，按最高频率值填补的结果和处理前差别最大。其余可视化图与此同理，在此不再赘述。