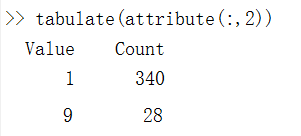
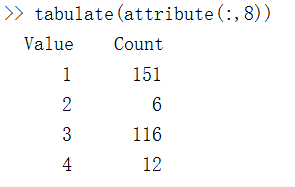
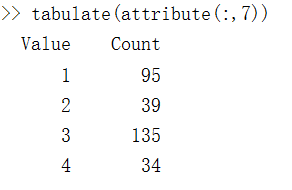
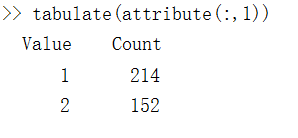
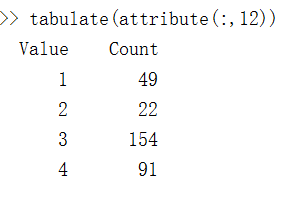
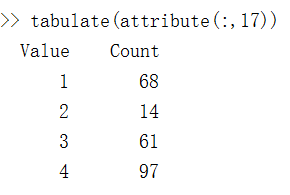
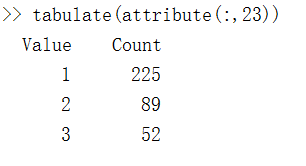
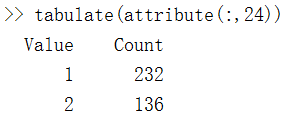
数据挖掘——马的疝病分析

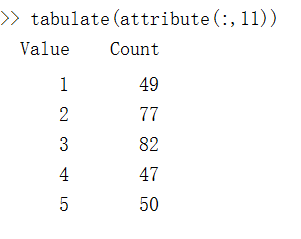
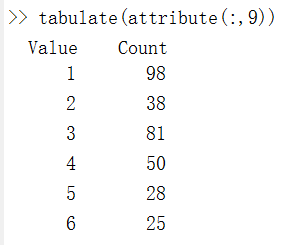
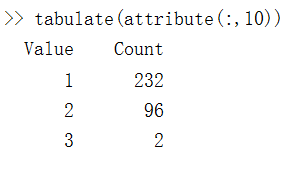
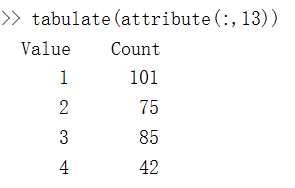
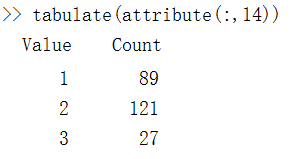
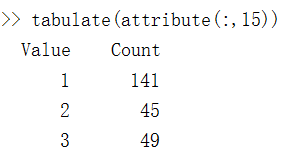
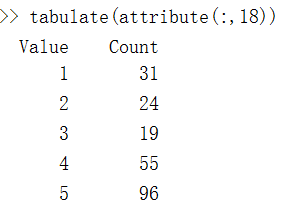
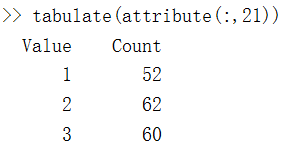
2120161061 王雨佳

1. 数据摘要

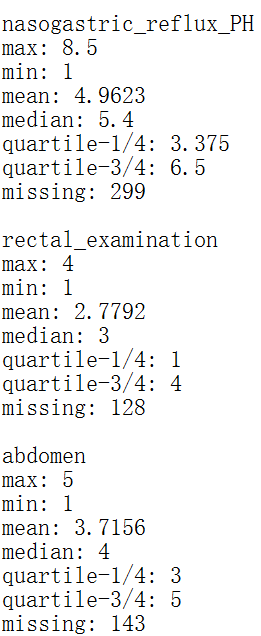
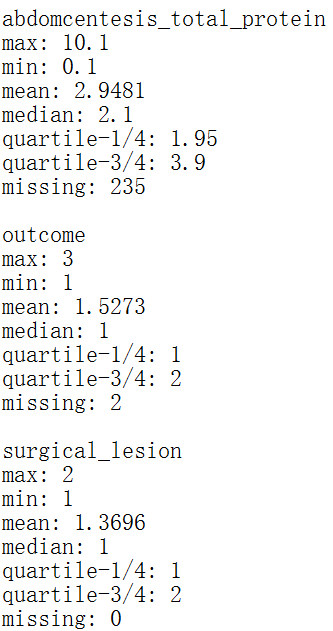
为方便处理，首先将数据中缺失数据“?”统一替换为“NaN”，且Hospital Number不具有意义，在处理过程中不考虑该属性。

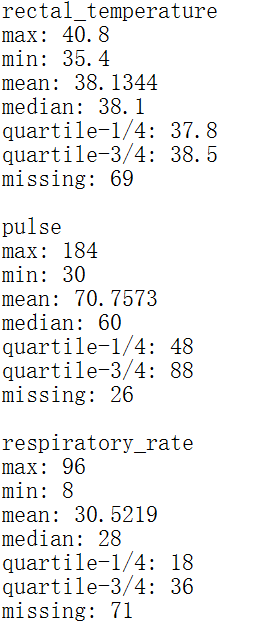
* 1. 对标称属性，给出每个可能取值的频数。

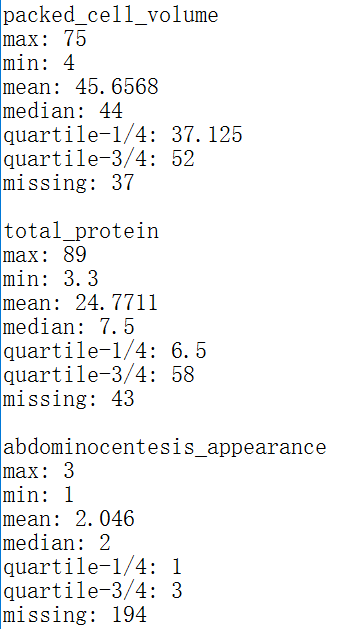
针对标称属性（surgery/Age/temperature of extremities/peripheral pulse / mucous membranes/capillary refill time/pain/peristalsis/abdominal distension/ nasogastric tube/nasogastric reflux/rectal examination/abdomen/ abdominocentesis appearance/outcome/surgical lesion/lesion1/lesion2/lesion3/cp\_data）每个可能取值的频数，利用matlab函数“tabulate”进行计算，由于属性较多，随机选择一部分属性进行结果展示，结果如下：



* 1. 数值属性（rectal temperature/pulse/respiratory rate/nasogastric reflux PH/ packed cell volume/total protein/abdomcentesis total protein），最大、最小、均值、中位数、四分位数及缺失值的个数如下图所示：





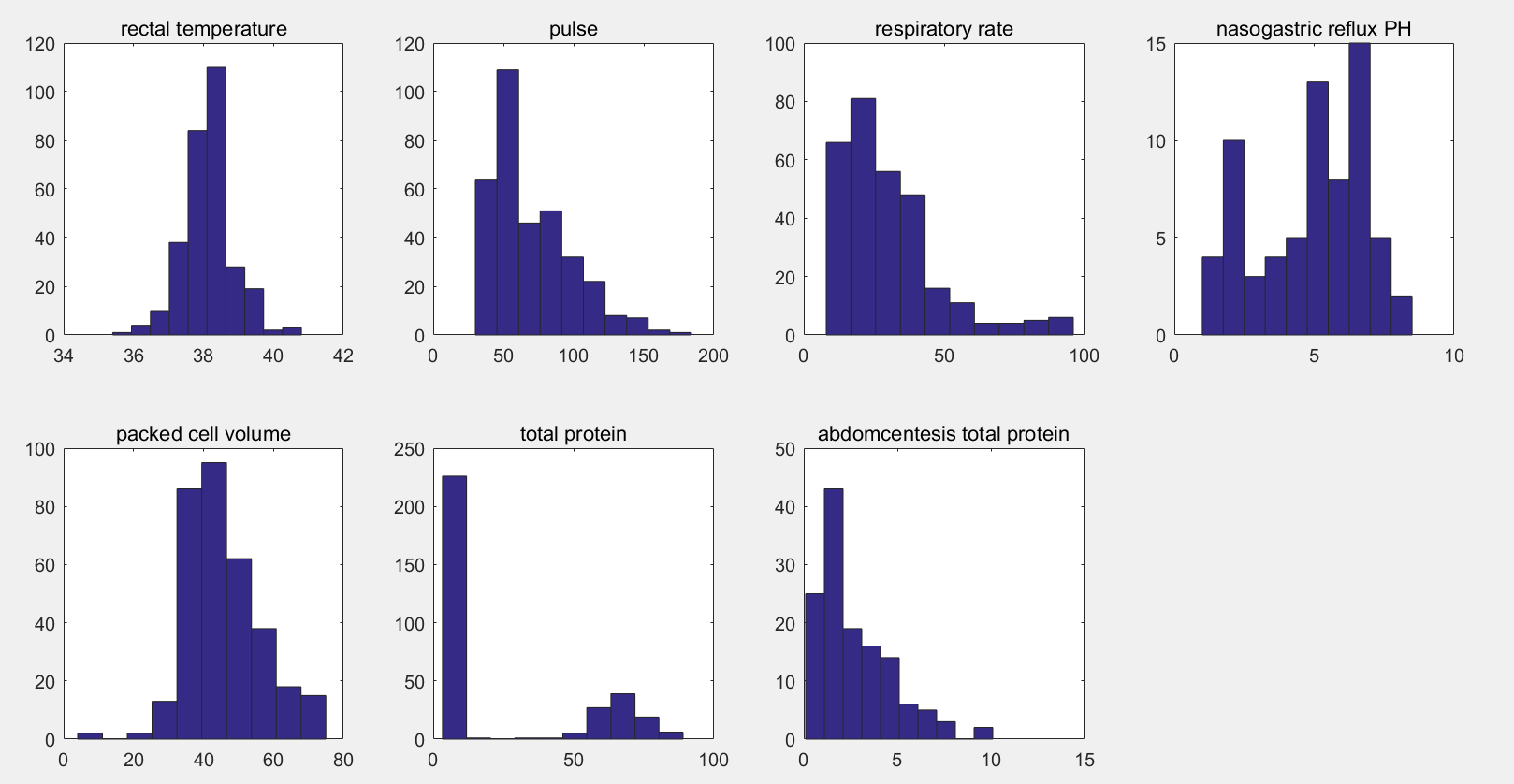


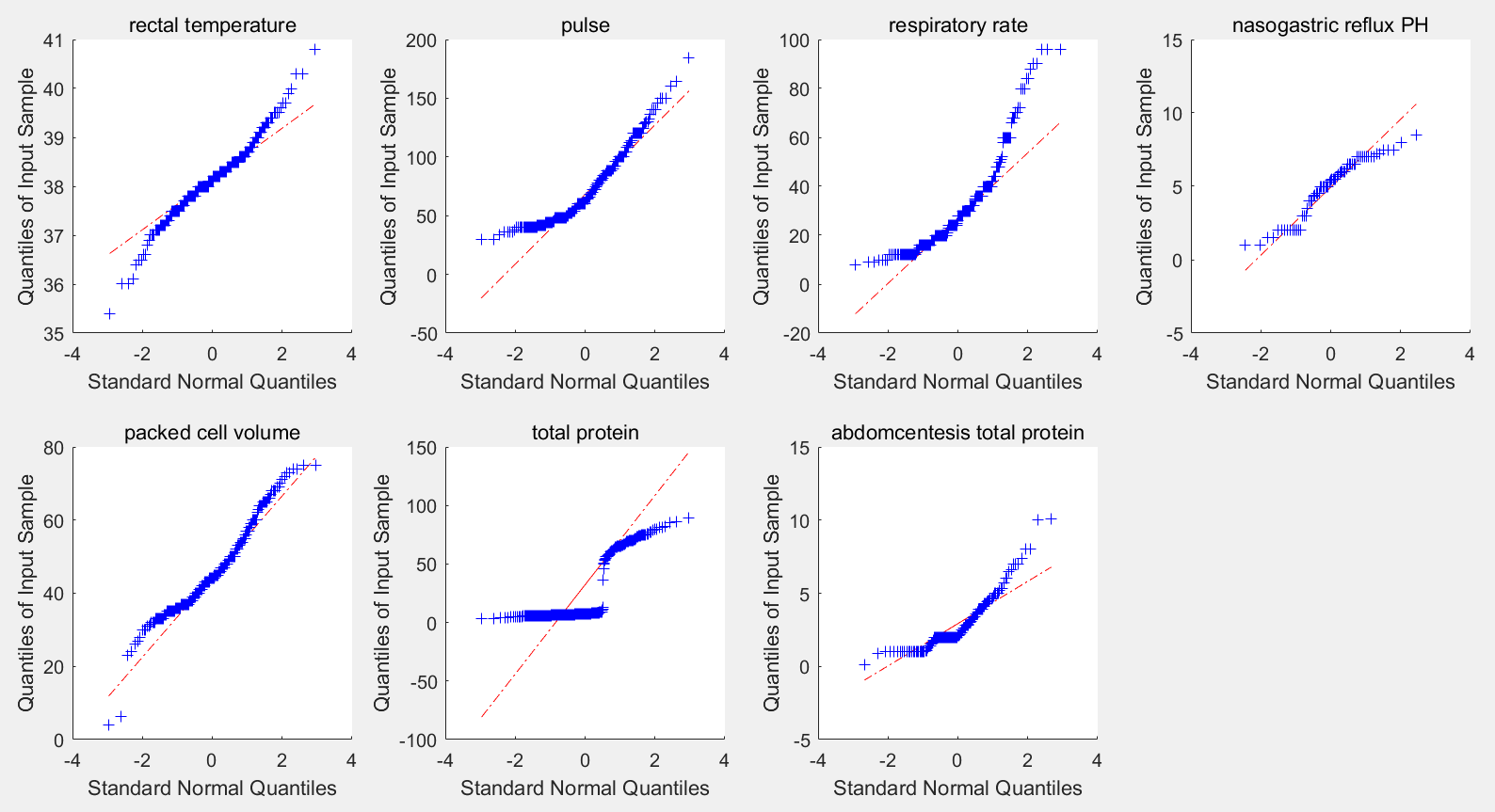
1. 数据可视化

针对数值属性，

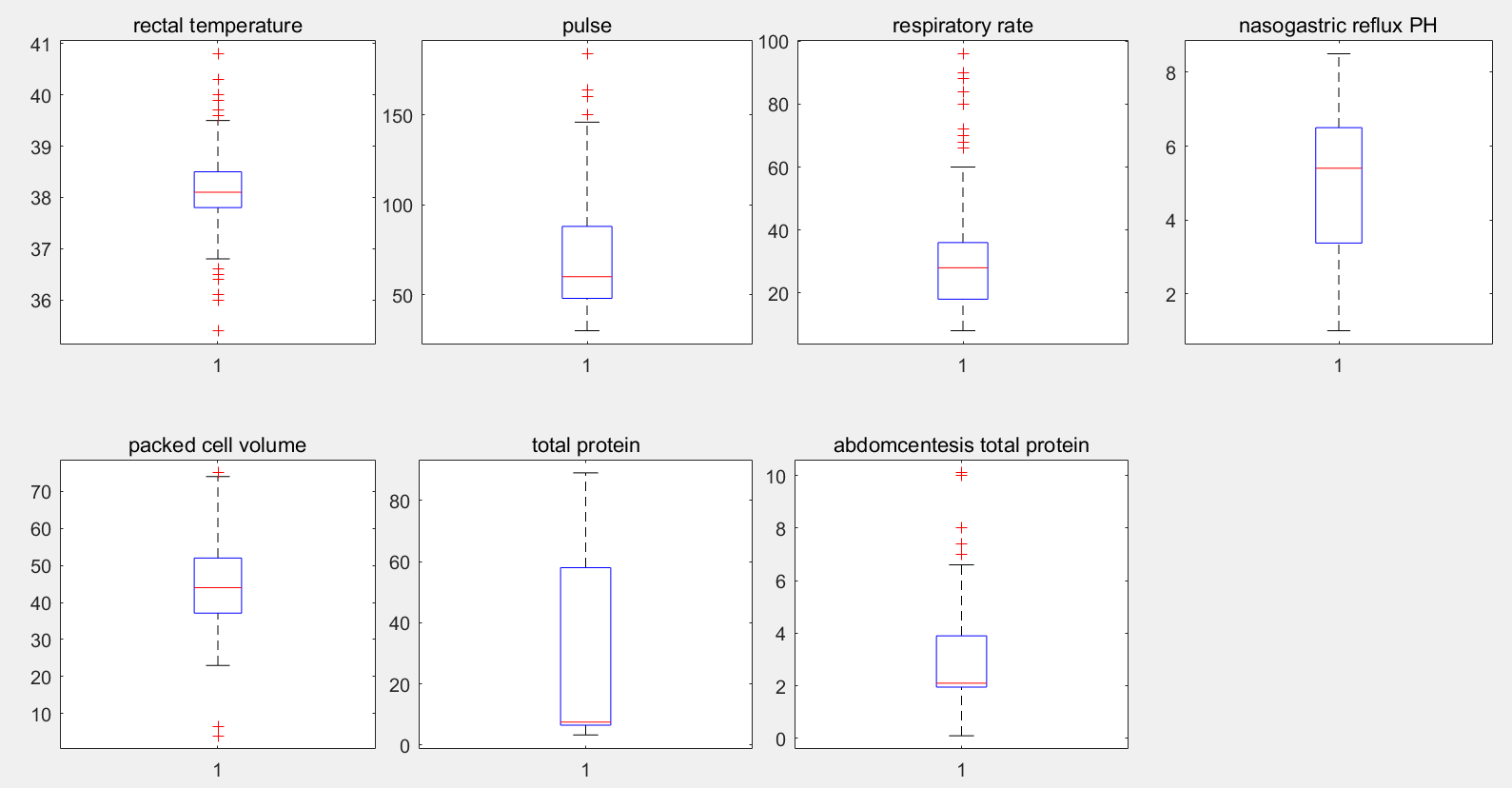
* 1. 绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。

针对各数值属性，直方图如下所示：



针对各数值属性，qq图如下图所示：

通过以上qq图，可以看出属性“rectal temperature”与“packed cell volume” 拟合正态分布较好。

* 1. 绘制盒图，对离群值进行识别。

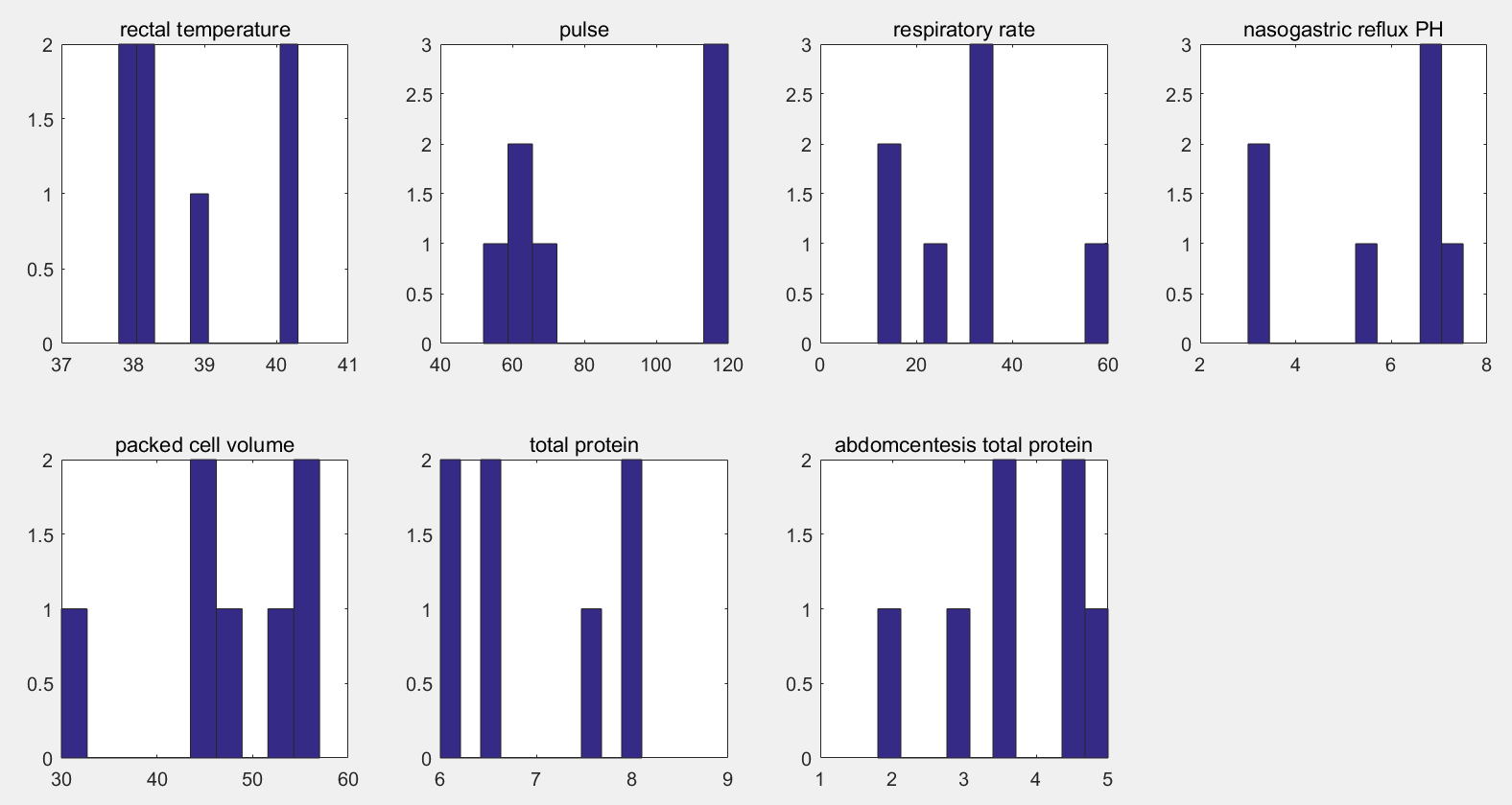
可以看出rectal temperature属性和respiratory rate属性相比较于其他属性的盒图具有较多的离群值。

1. 数据缺失的处理

数据集中有30%的值是缺失的，因此需要先处理数据中的缺失值。

分别使用下列四种策略对缺失值进行处理：

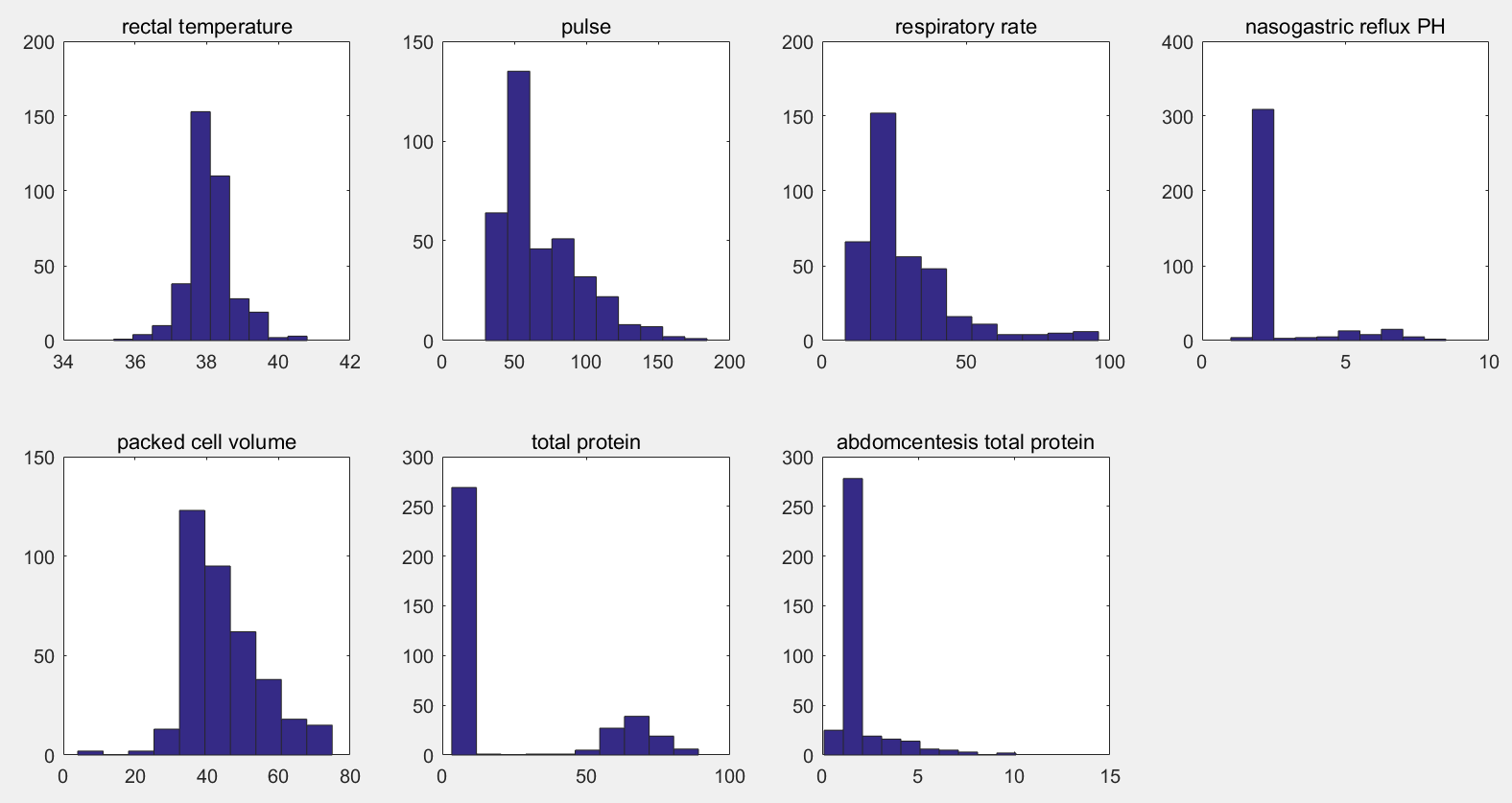
* 剔除缺失数据
* 按最高频率值：对于某一属性下的缺失，使用该属性的最高频率值代替缺失的数据；
* 按属性填补：计算两个属性的相关性，相关性越大表明可以根据另一属性推断缺失属性的值。通过另一属性的回归分析，计算当前的缺失值；
* 按相似性填补：计算两个样本的相似程度，越相似证明越可以使用该样本推断当前含缺失值的样本。

1. 将缺失部分剔除

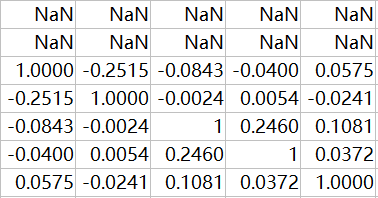
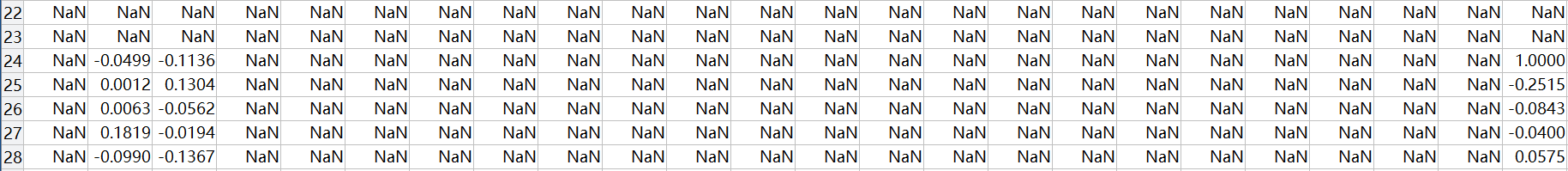
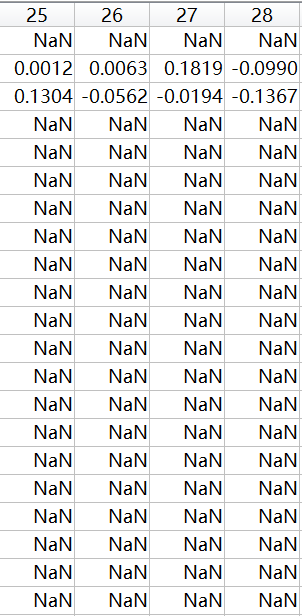
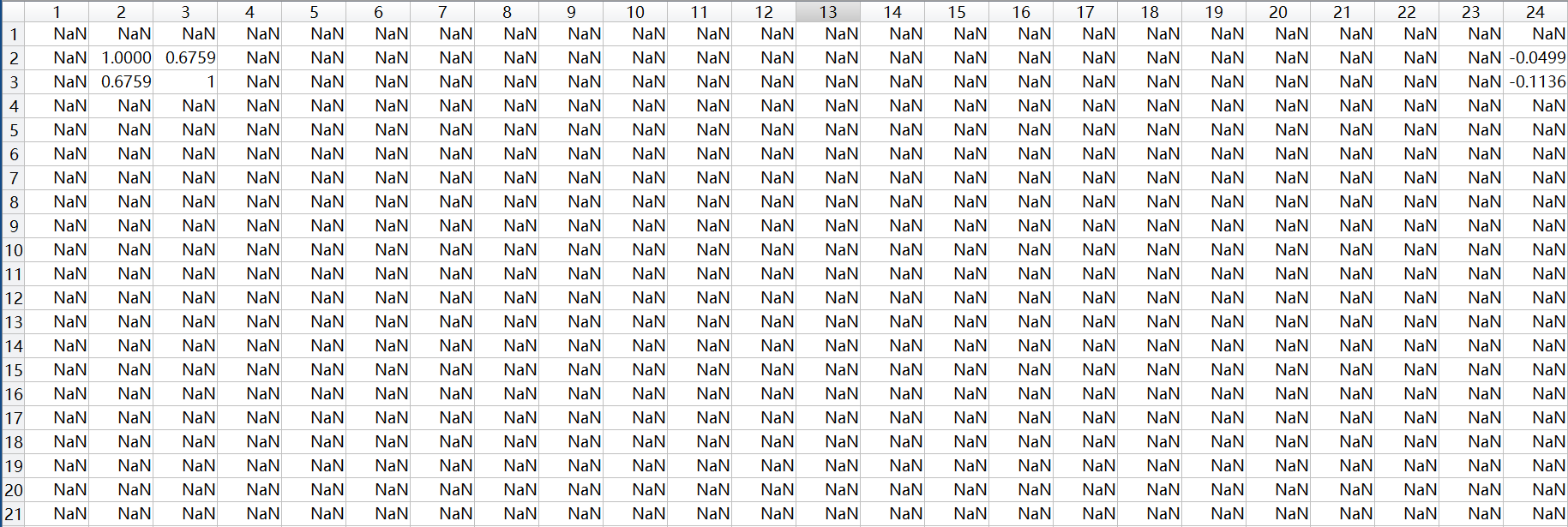
经过删除，该数据集仅剩7条数据，用词方法对缺失值进行处理效果并不好。

1. 用最高频率值来填补缺失值

根据直方图，用最高频率值填补缺失值后，变化不明显。



1. 通过属性的相关关系来填补缺失值



有上述结果可以看出，数值属性之间的相关性均不高。

1. 通过数据对象之间的相似性来填补缺失值

数据相似性结果见“attribute\_similarity.txt”,根据数据对象之间的相似性进行填补，结果如下所示：

