数据挖掘——马的疝病分析

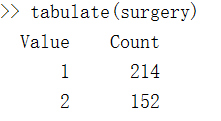
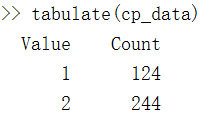
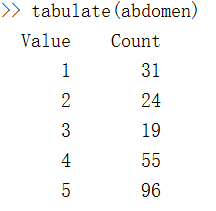
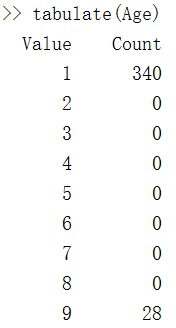
2120161061 王雨佳

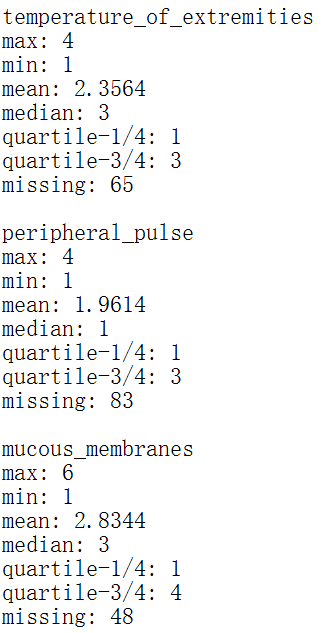
1. 数据摘要

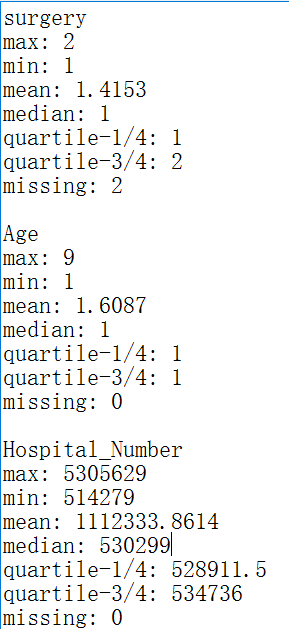
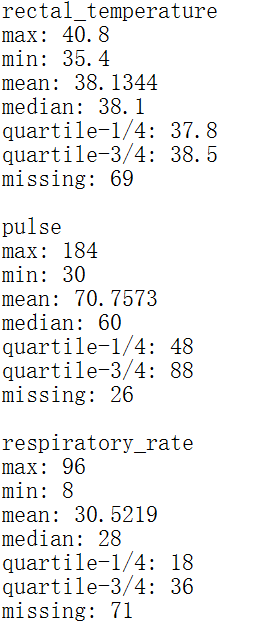
为方便处理，首先将数据中确实数据“?”统一替换为“NaN”。

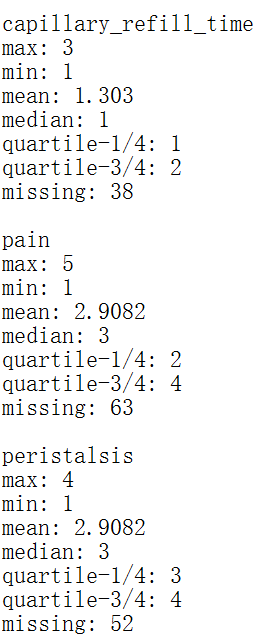
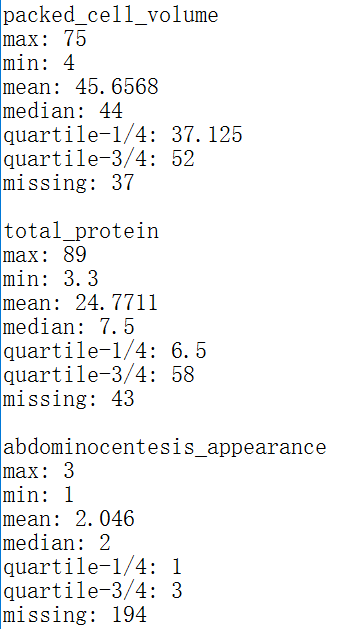
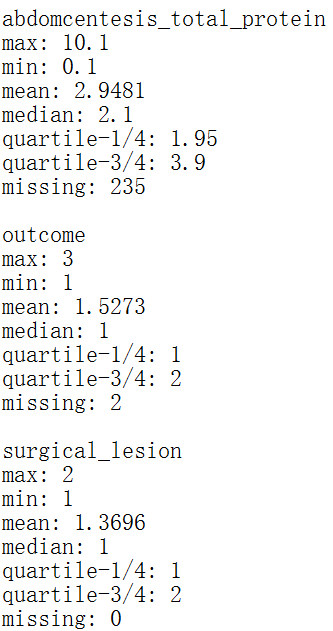
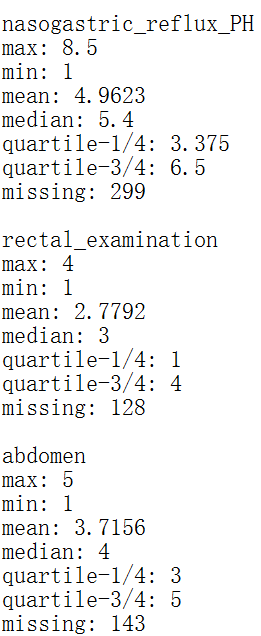
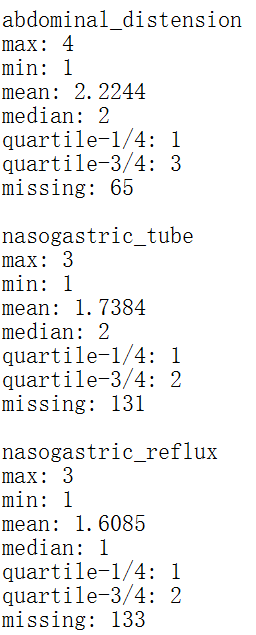
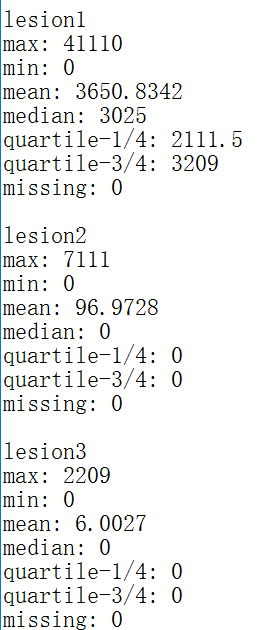
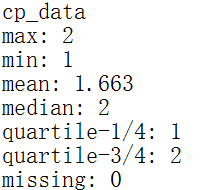
* 1. 对标称属性，给出每个可能取值的频数。

针对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，随机选择几个属性（surgery/Age/abdomen/cp\_data）进行结果展示，结果如下：



* 1. 数值属性，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：

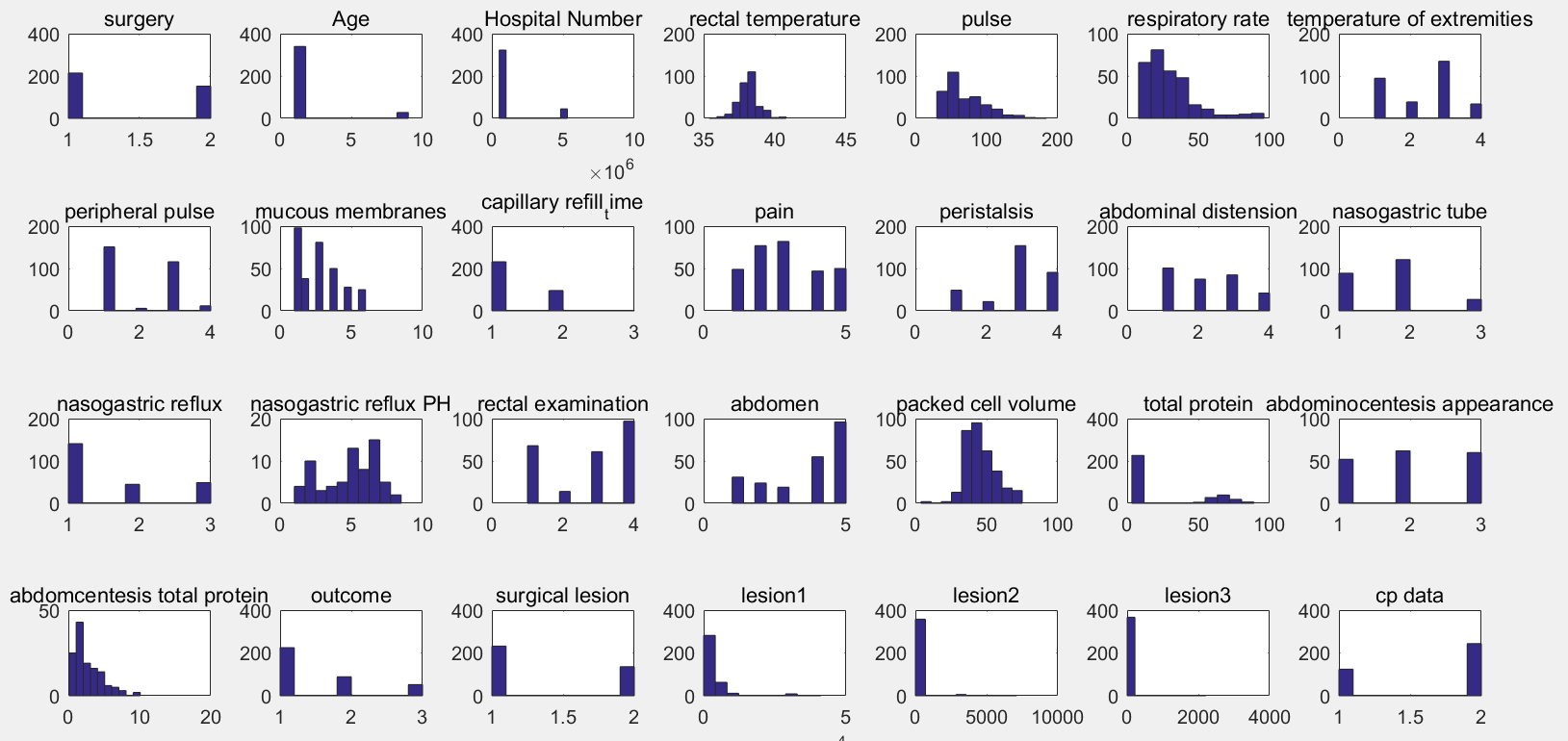


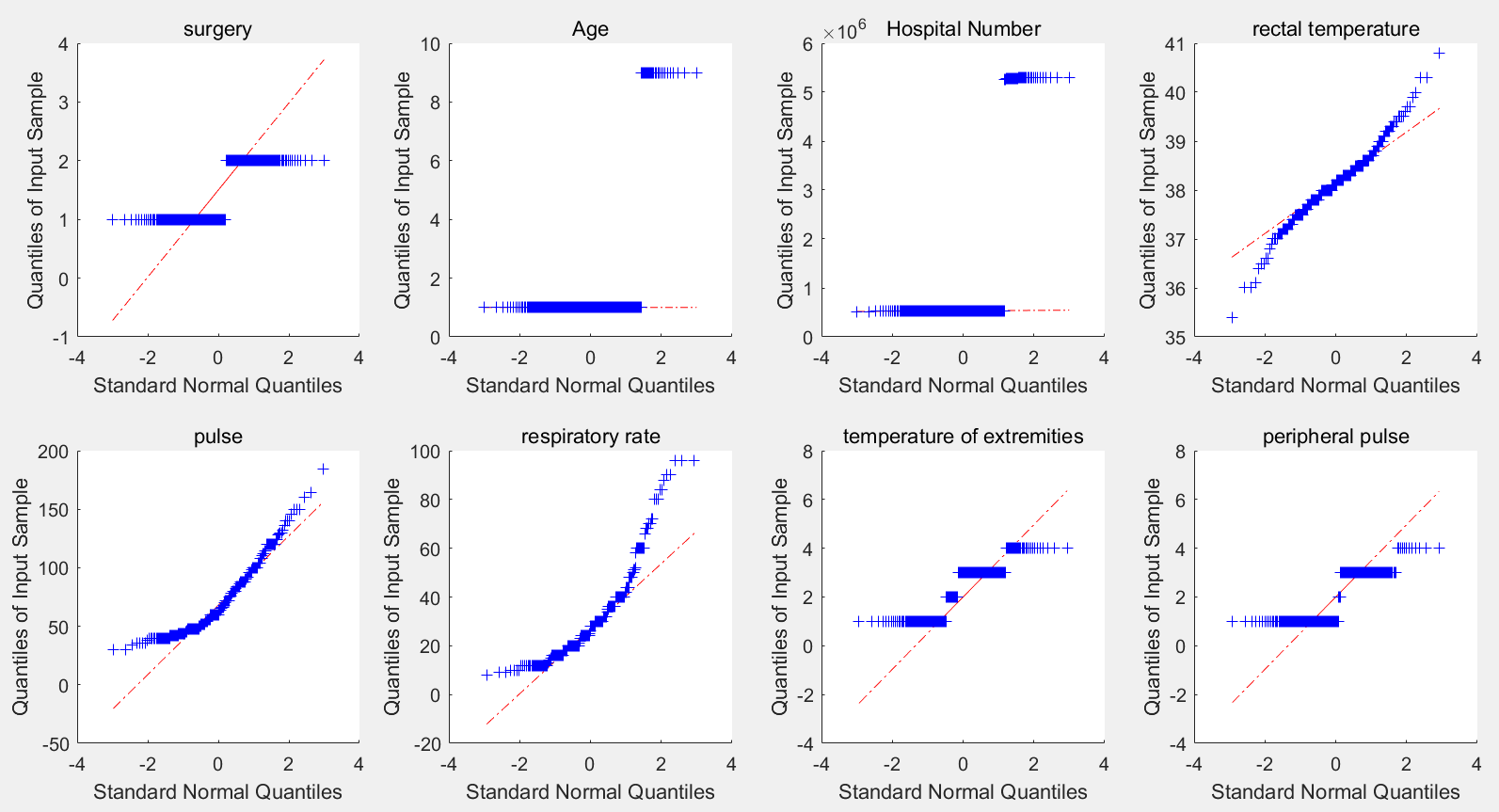


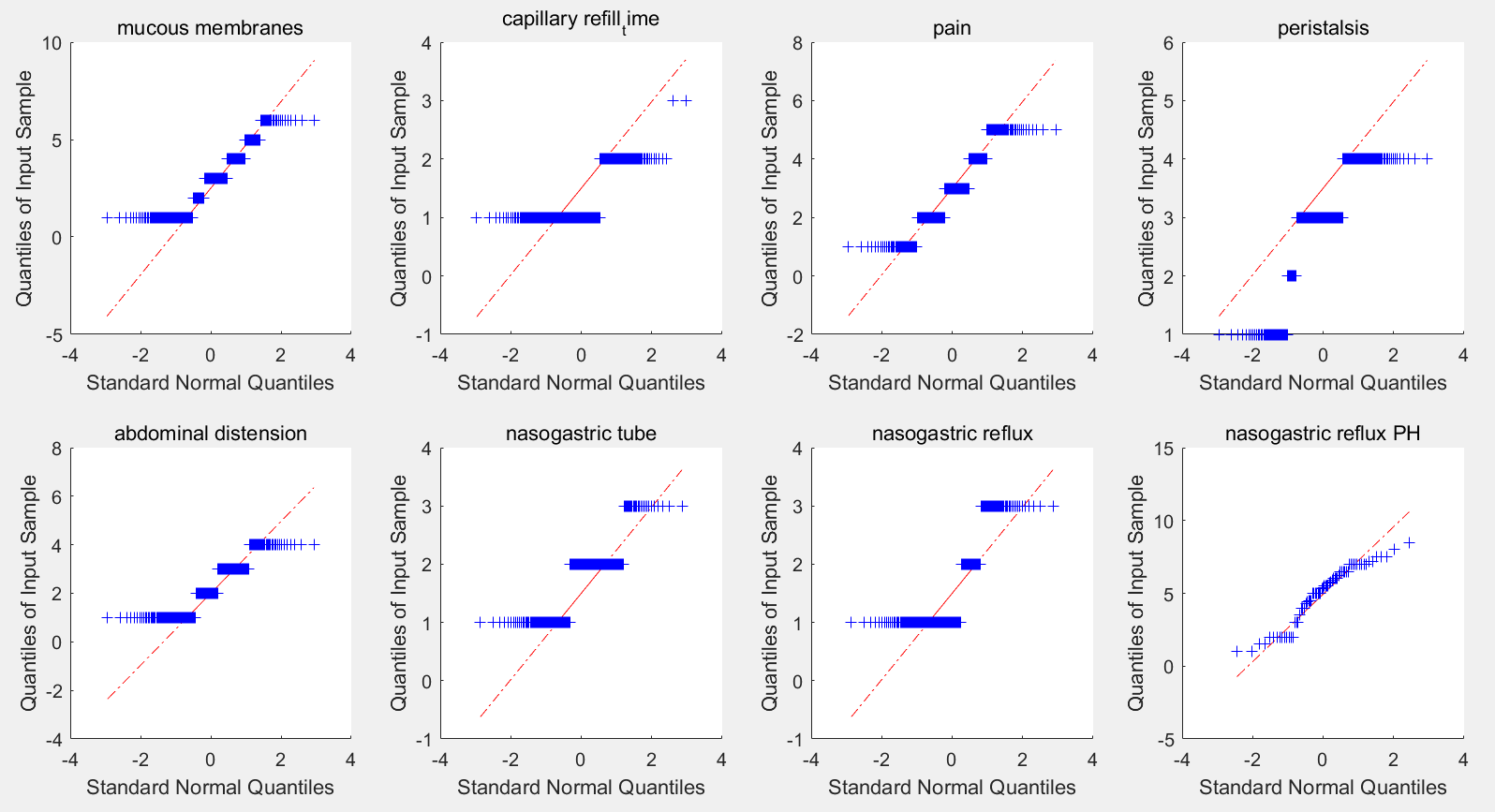
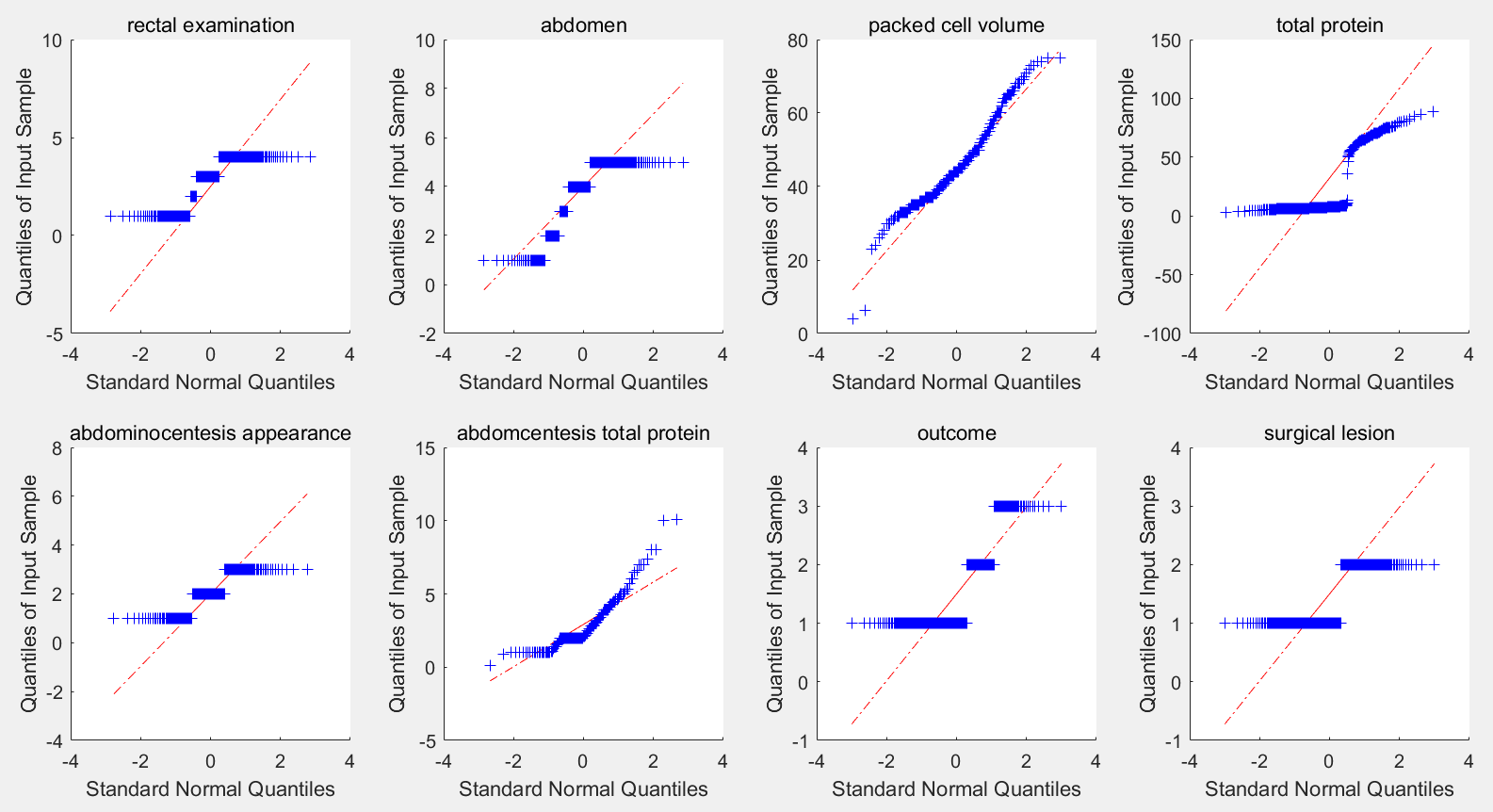
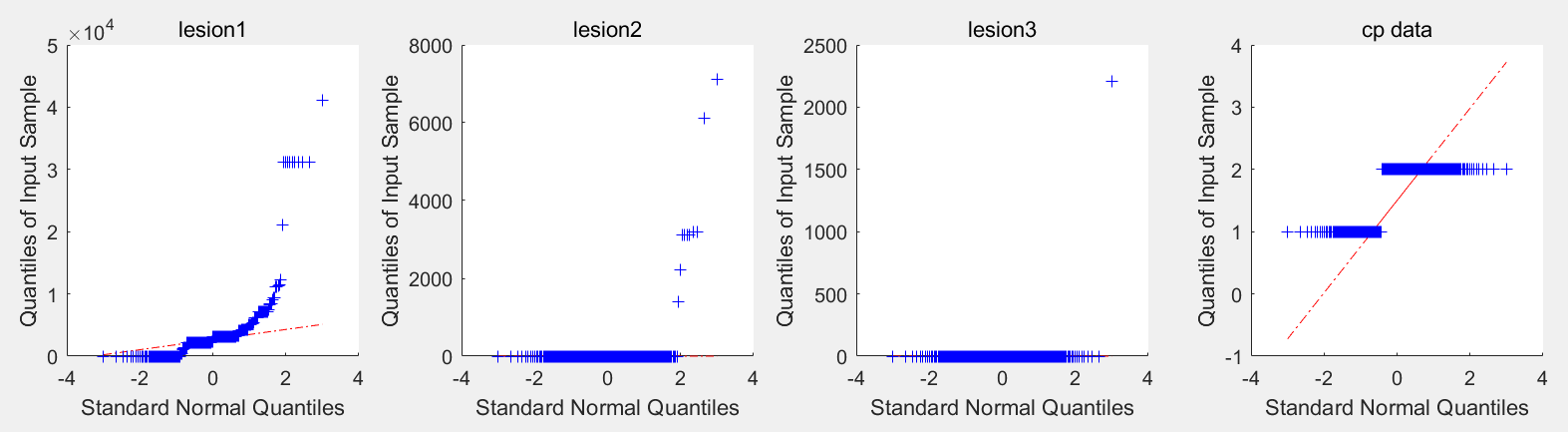
1. 数据可视化

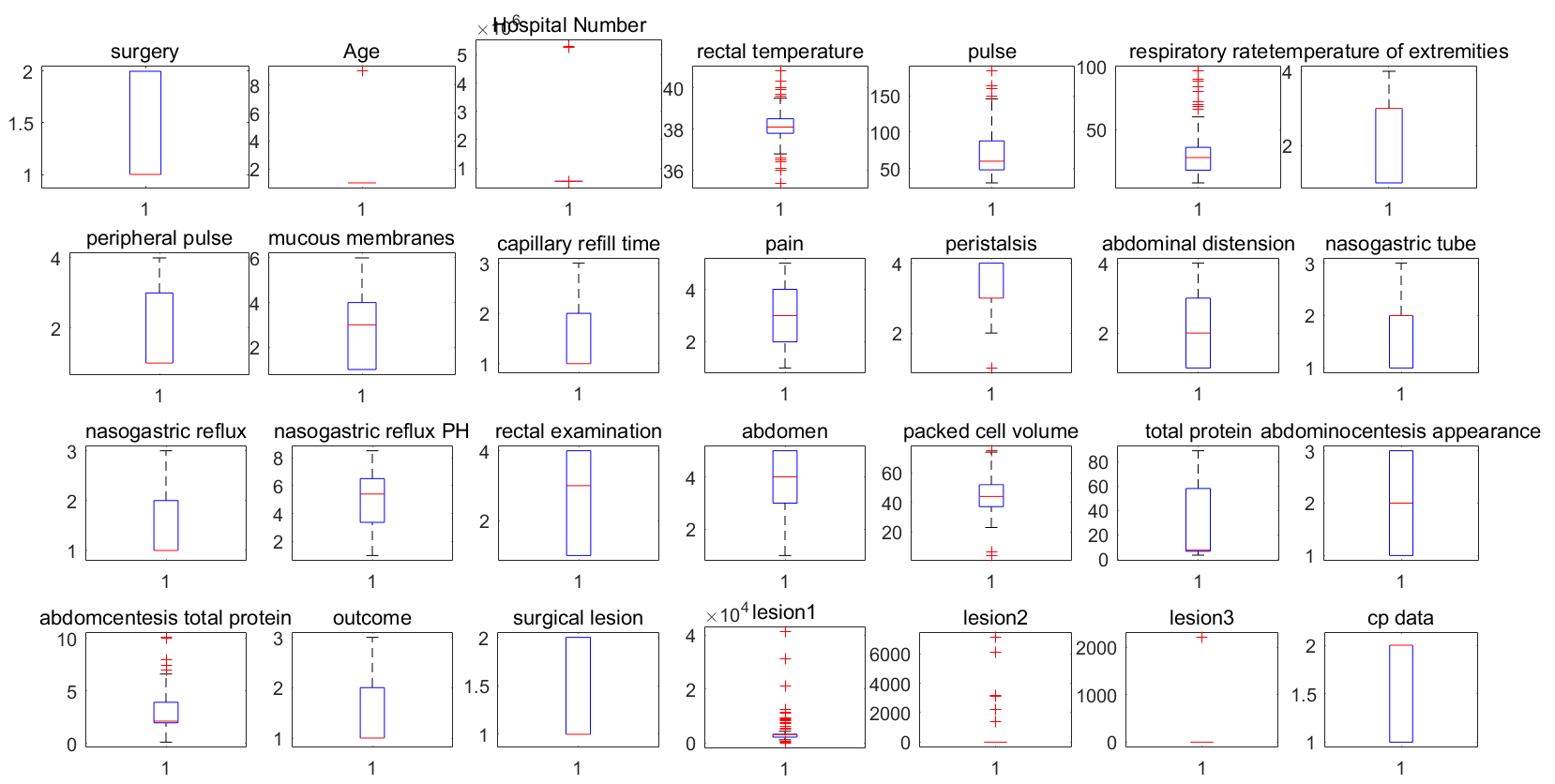
针对数值属性，

* 1. 绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。

针对各标称属性，直方图如下所示：

针对各标称属性，qq图如下图所示：

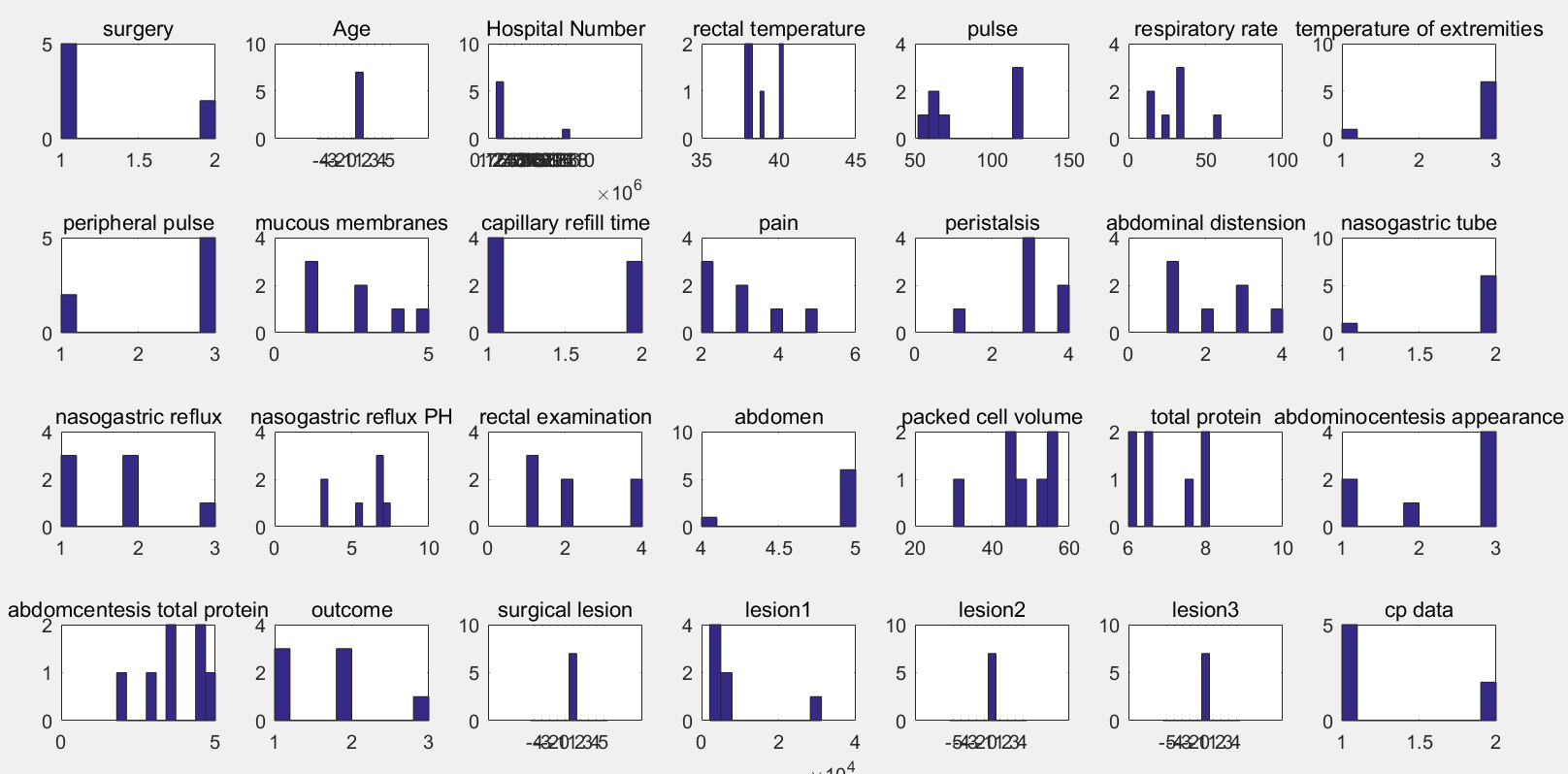
 通过以上qq图，可以看出属性“packed cell volume”服从正态分布。

* 1. 绘制盒图，对离群值进行识别。

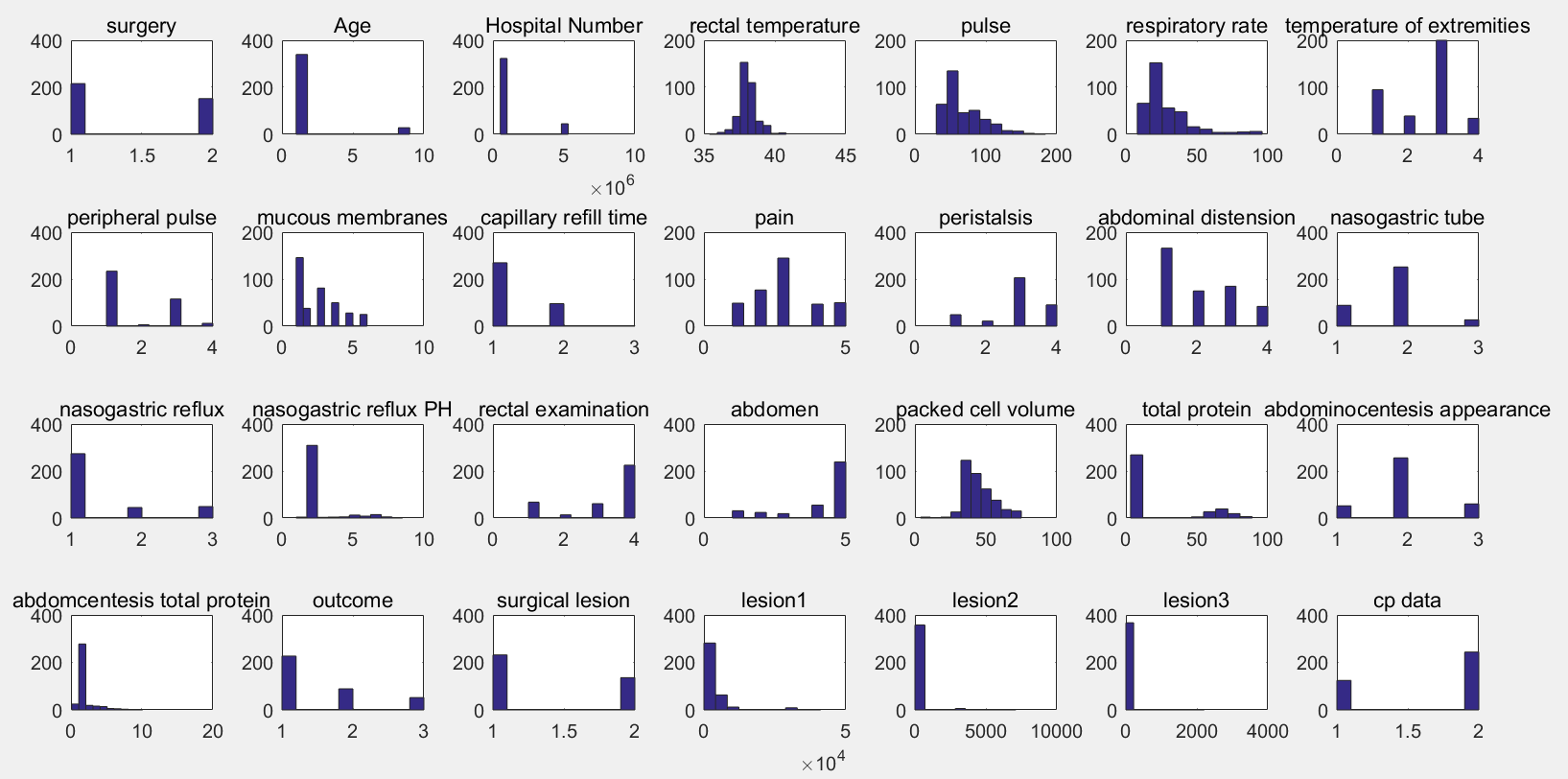
1. 数据缺失的处理

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：

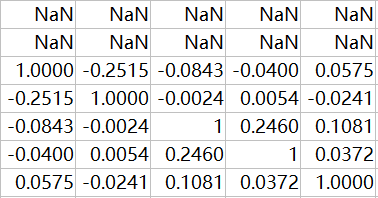
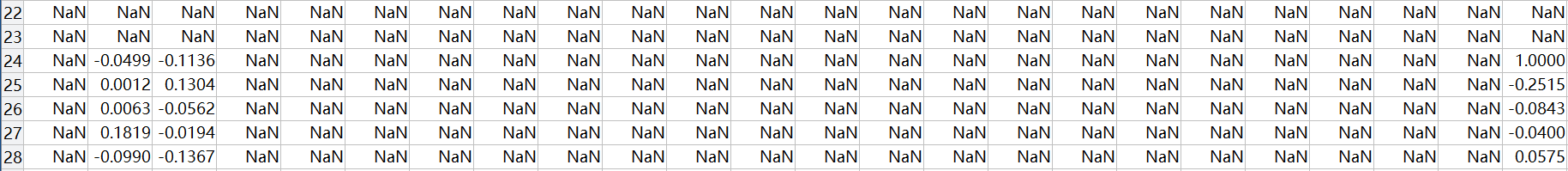
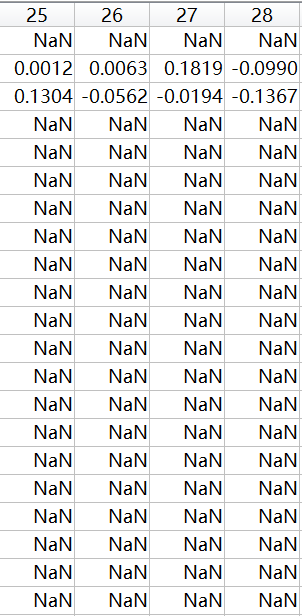
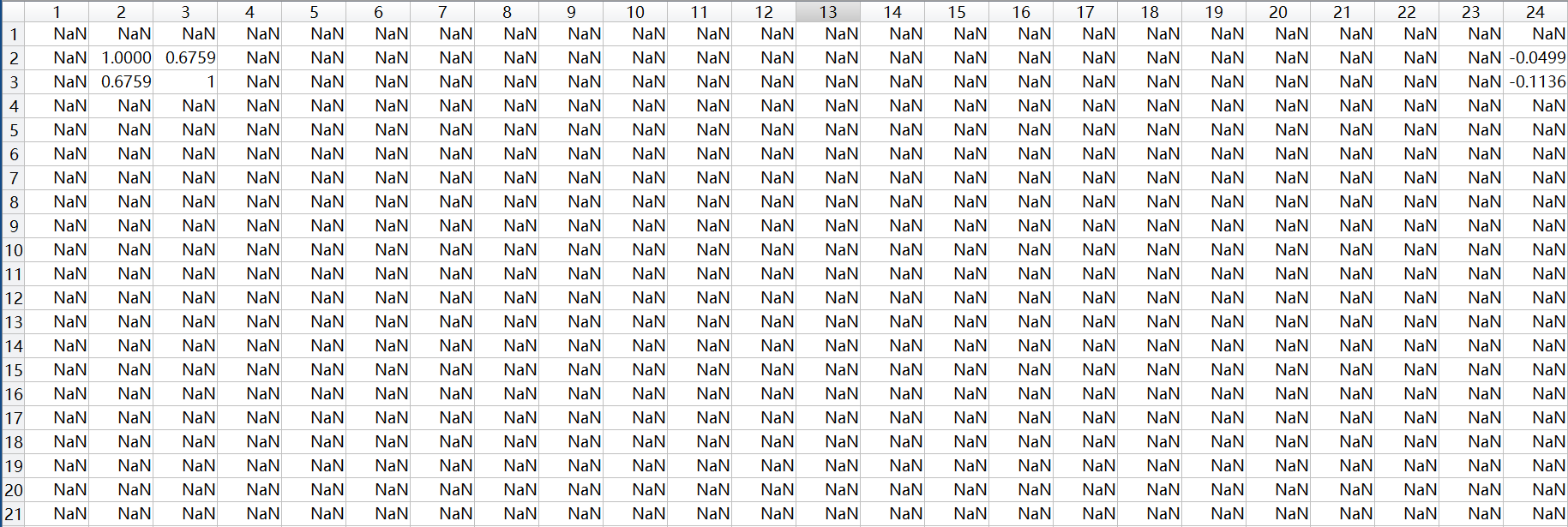
1. 将缺失部分剔除

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。

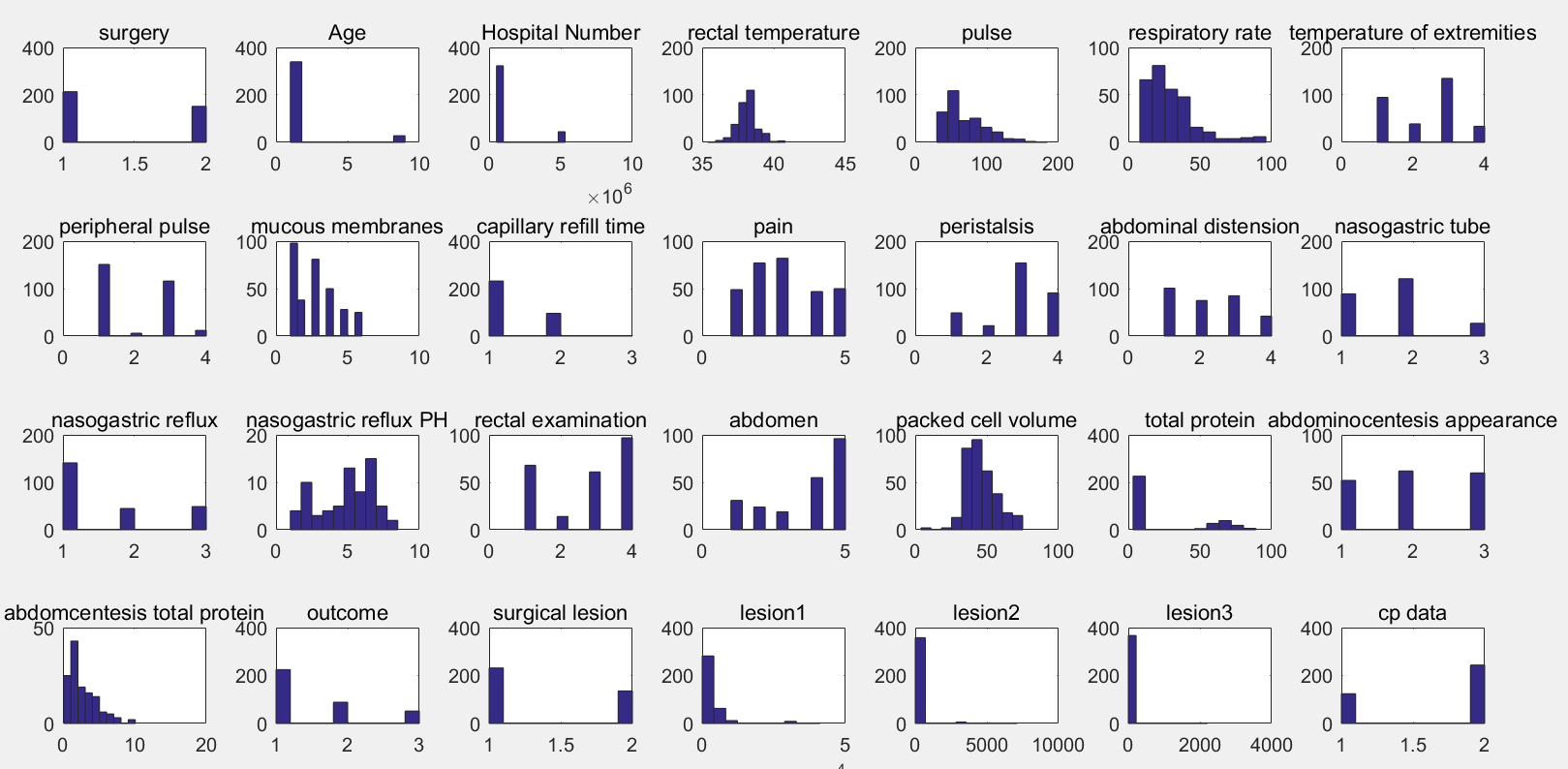
1. 通过属性的相关关系来填补缺失值



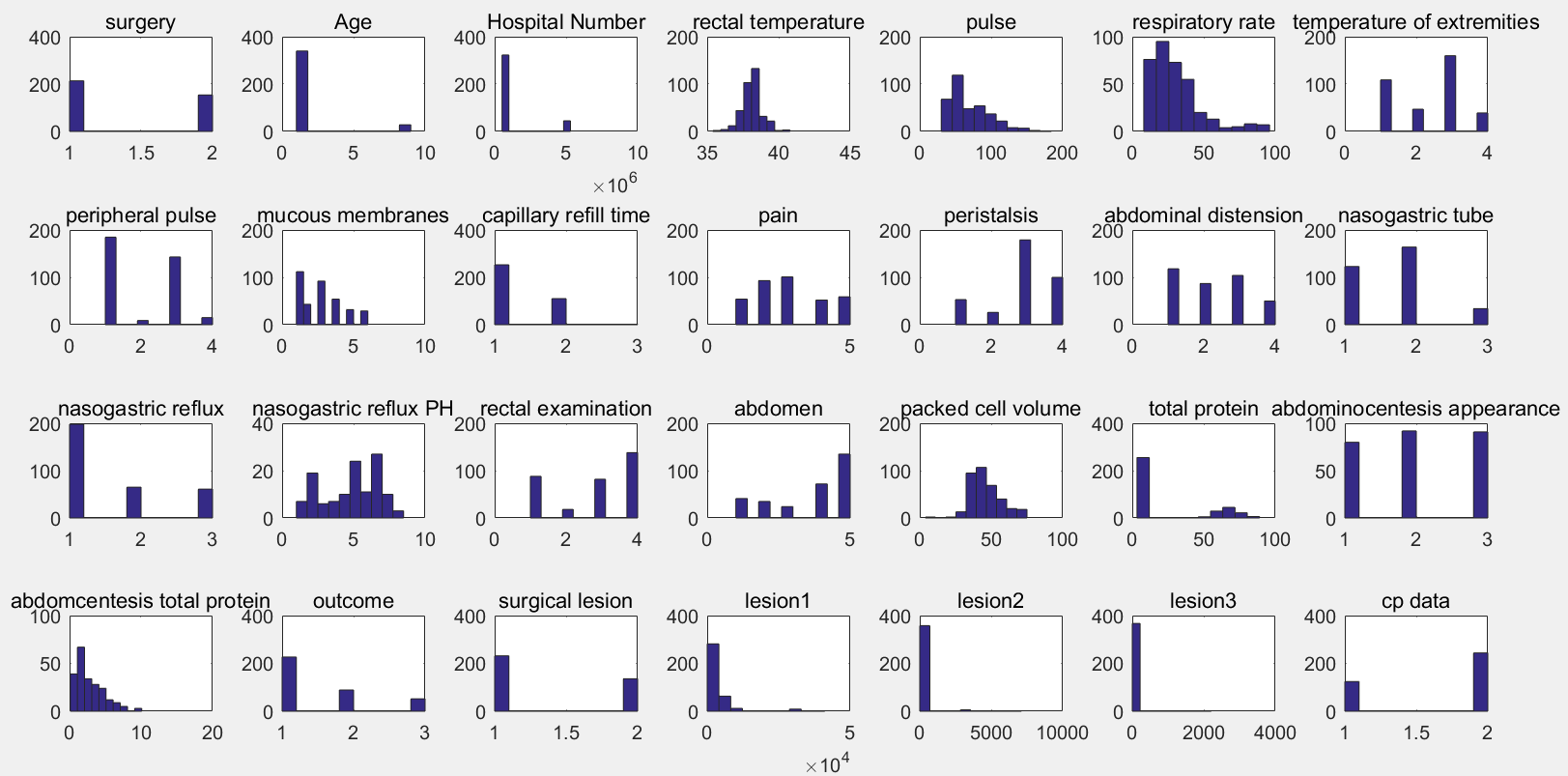
有上述结果可以看出，属性“Age”与“Hospital Number”显著相关（0.6759），“surgical lesion”与“lesion1”微相关（-0.2515），“lesion2”与“lesion3”微相关（0.2460）。

Hospital Number = +(4.9647e+05) \* Age + (3.1367e+05)

Age = +(9.2019e-07) \* Hospital Number + (0.58514)



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：