**马的疝病分析**

1. **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。

1. **数据说明**

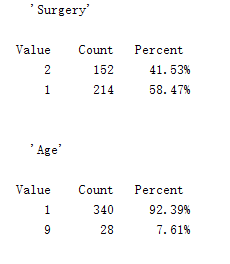
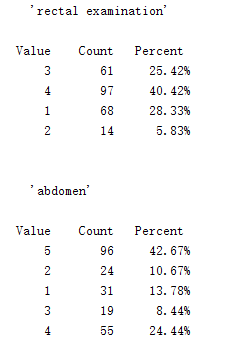
共有368个样本包含28个属性，其中有30%的缺失值。

1. **数据分析**

环境: matlab2016a, win10

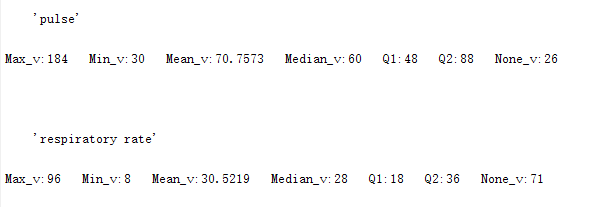
* 1. **数据摘要**

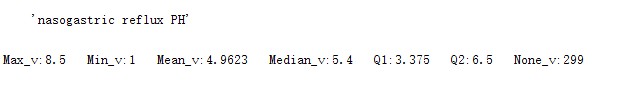
1. 对标称属性，给出每个可能取值的频数。结果如图：

四个属性的相关频数统计如上图，程序分别统计了不同属性所有可能的取值（value）以及出现的次数(count)和所占的百分比(percent)。

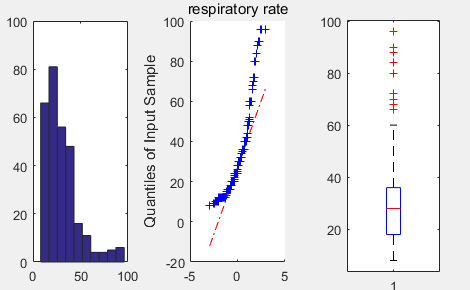
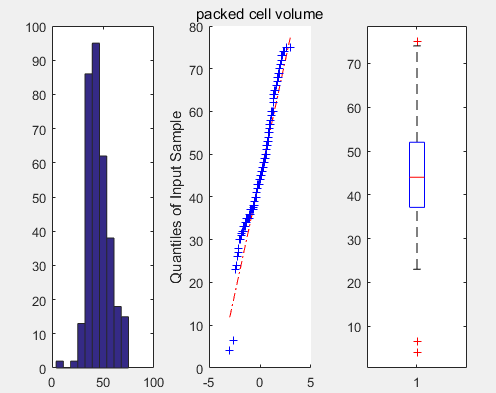
1. 数值属性，给出最大、最小、均值、中位数、四分位数以及缺失值的个数。结果如图：

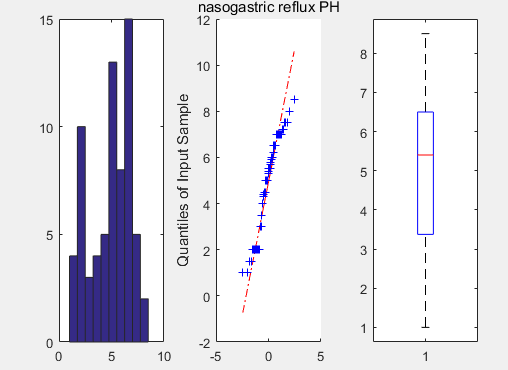
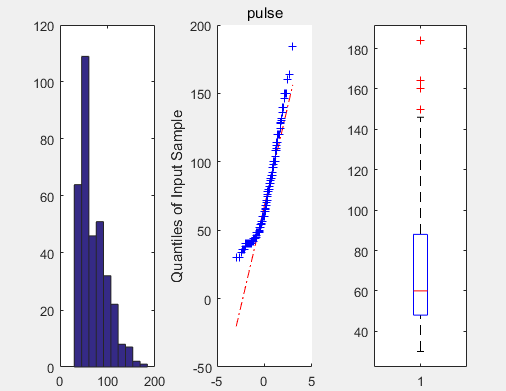


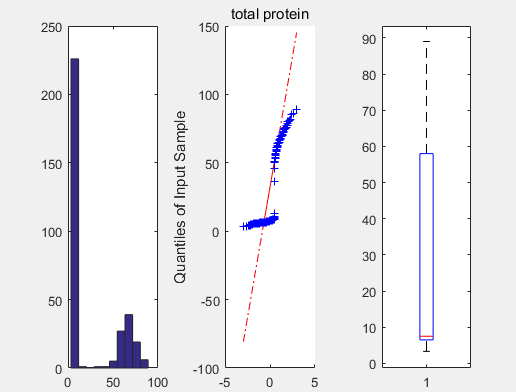
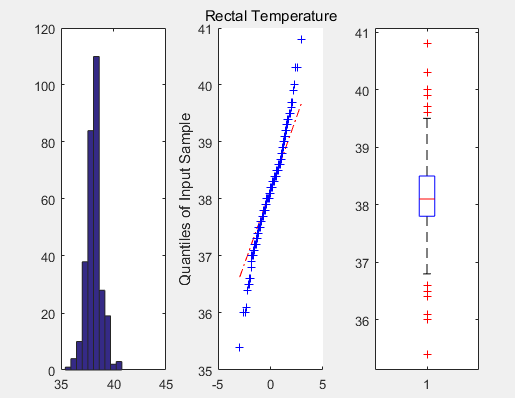


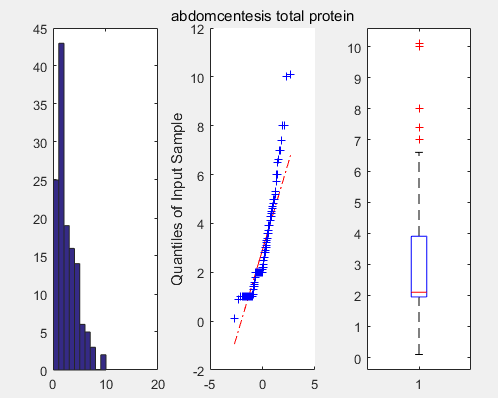
* 1. **数据可视化**

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。2、绘制盒图，对离群值进行识别。可视化如下图：



左边为直方图，中间为qq图，右边为盒图。可以看出左边图中越接近正态分布，qq图就越贴合中间的直线。比如属性'packed cell volume'、'Rectal Temperature'就比较符合正态分布，而属性'total protein'就很不符合正态分布。

然后看右边的盒图，位于盒图中间的红线为中位数，盒顶端和低端为四分位数，盒两端的红线为最大、最小值，加号为离群点，可以看到属性'respiratory rate'、'Rectal Temperature'有较多的离群点，而属性'nasogastric reflux PH'、'total protein'则有较少的离群点。

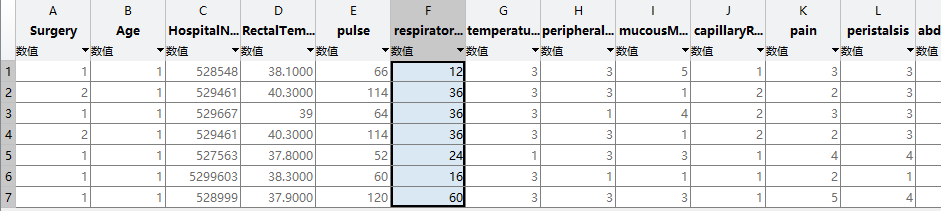
1. **数据缺失值的处理**

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。

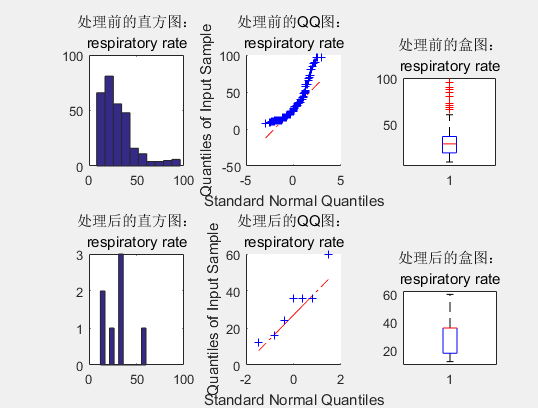
分别使用下列四种策略对缺失值进行处理可视化的对比新旧数据集（以属性'respiratory rate'为例，其余属性的可视化情况与其类似）:

1. **将缺失部分剔除**

剔除后的数据形式：



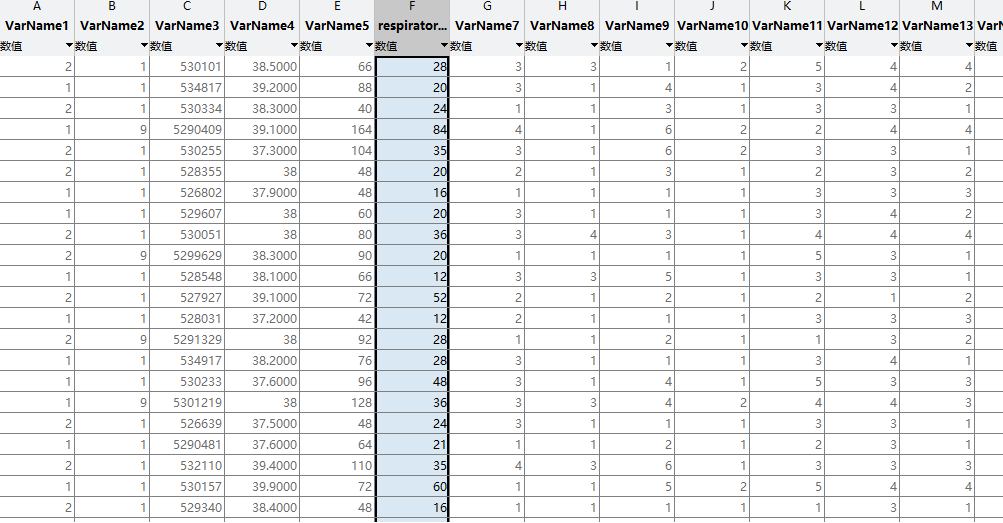
新旧数据集可视化对比：



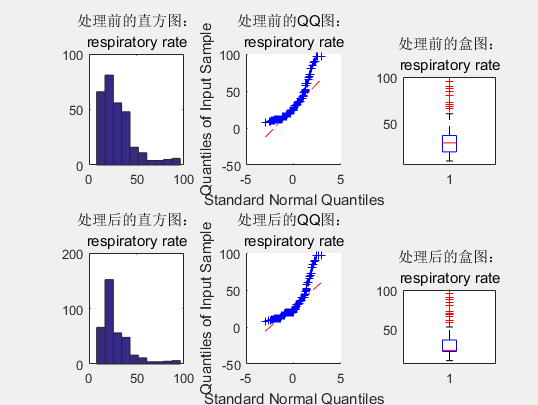
总结：可以看到，由于各个属性的存在缺失值的样本不同，导致直接删除缺失值样本之后样本数量所剩无几，所以这种处理缺失值的方法代价太大不可取。

1. **用最高频率值来填补缺失值。**

处理后的数据形式：

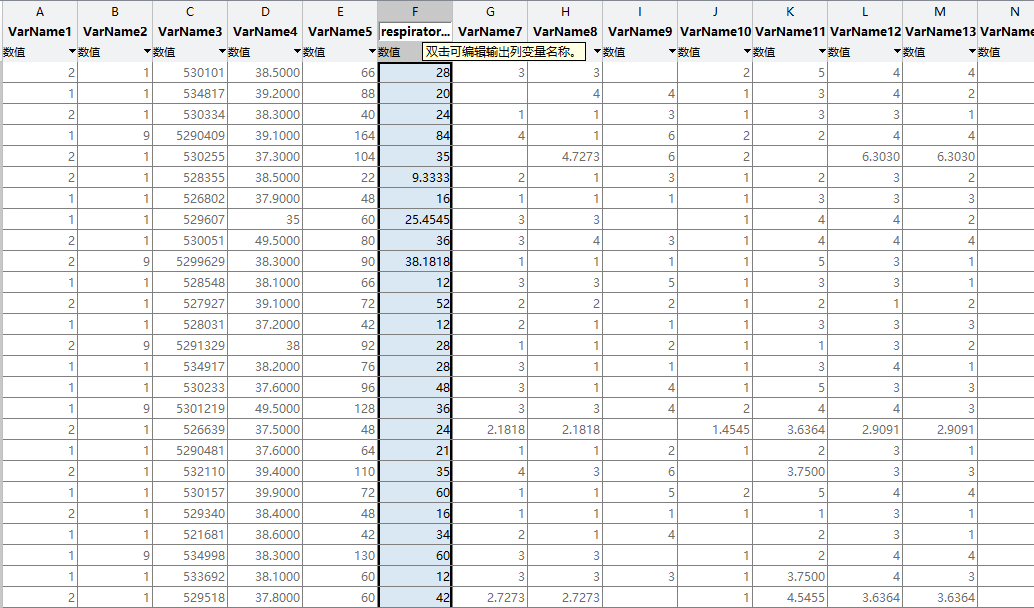


新旧数据可视化对比：

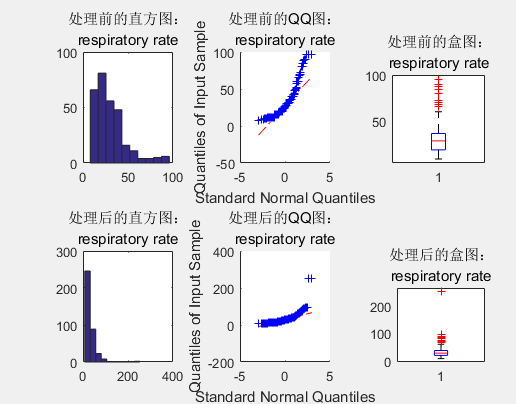


可以看到该方法处理缺失值之后数据分布基本符合原数据，但是会导致众数规模更大。

1. **通过属性的相关关系来填补缺失值**



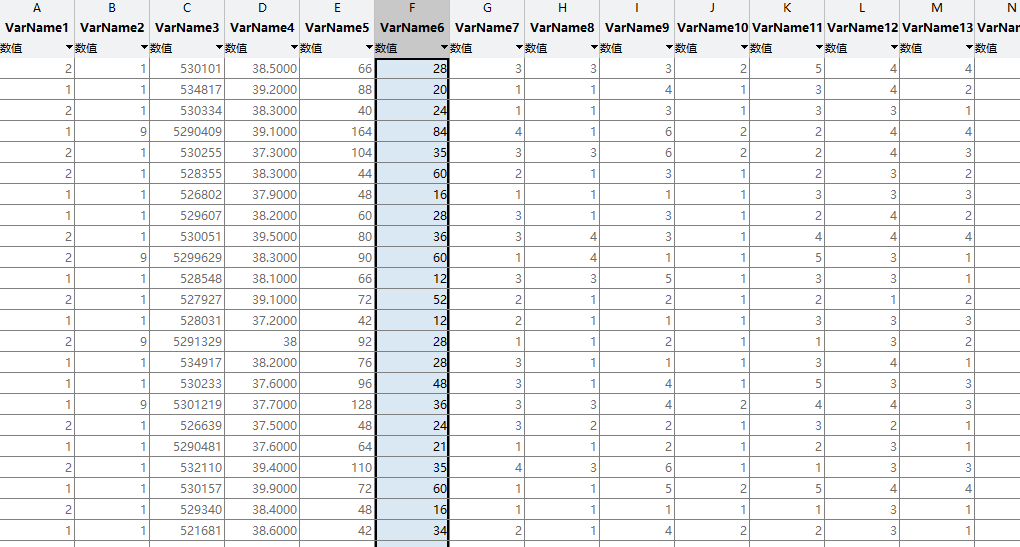
新旧数据集可视化对比：



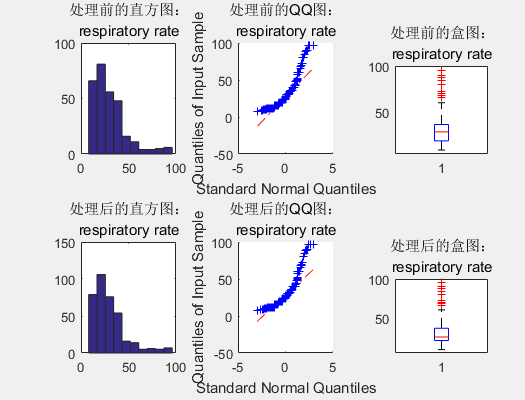
通过此方法处理数据使得数据分布比原先更加密集。

1. **通过数据对象之间的相似性来填补缺失值**

处理后的数据:



新旧数据可视化对比:



通过该方法处理缺失值最为符合原数据的形式。

*备注：具体文档代码详见README.TXT*