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。

2. 数据说明

**下载数据:** [地址](http://archive.ics.uci.edu/ml/datasets/Horse+Colic)

共368个样本，28个特征。关于特征的详细说明见下载链接。

每条记录由28个变量构成，21个是标称变量，分别是第1、2、3、7、8、9、10、11、12、13、14、15、17、18、21、23、24、25、26、27、28个属性；7个是数值数据，分别是第4、5、6、16、19、20、22个属性。

3. 数据分析要求

3.1 数据可视化和摘要

数据摘要：

* 对标称属性，给出每个可能取值的频数，
* 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。

数据的可视化：

针对数值属性，

* 绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。
* 绘制盒图，对离群值进行识别

3.2 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:

* 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

处理后，可视化地对比新旧数据集。

4. 提交内容

* 分析报告
* 分析程序
* 预处理后的数据集

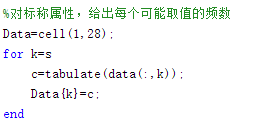
5.实验过程

1.加载所需库文件：

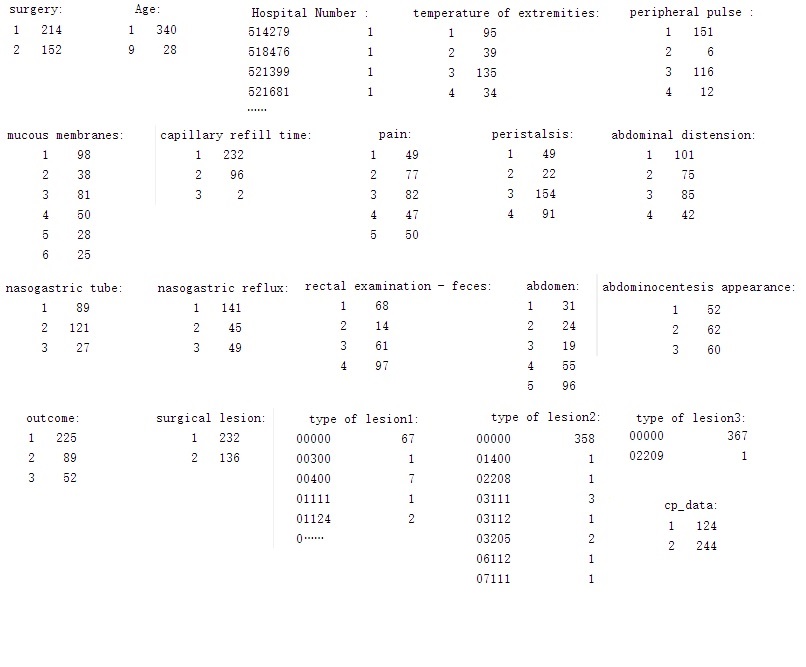


2.读取实验数据并获取数据摘要：

(1).处理标称属性数据：

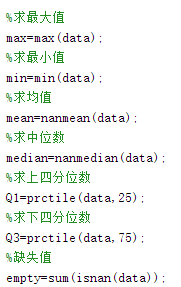


结果如下：



其中Hospital Number属性和type of lesion1属性的可能取值过多，就不一一列举。

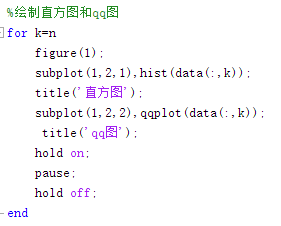
(2).处理数值属性数据：



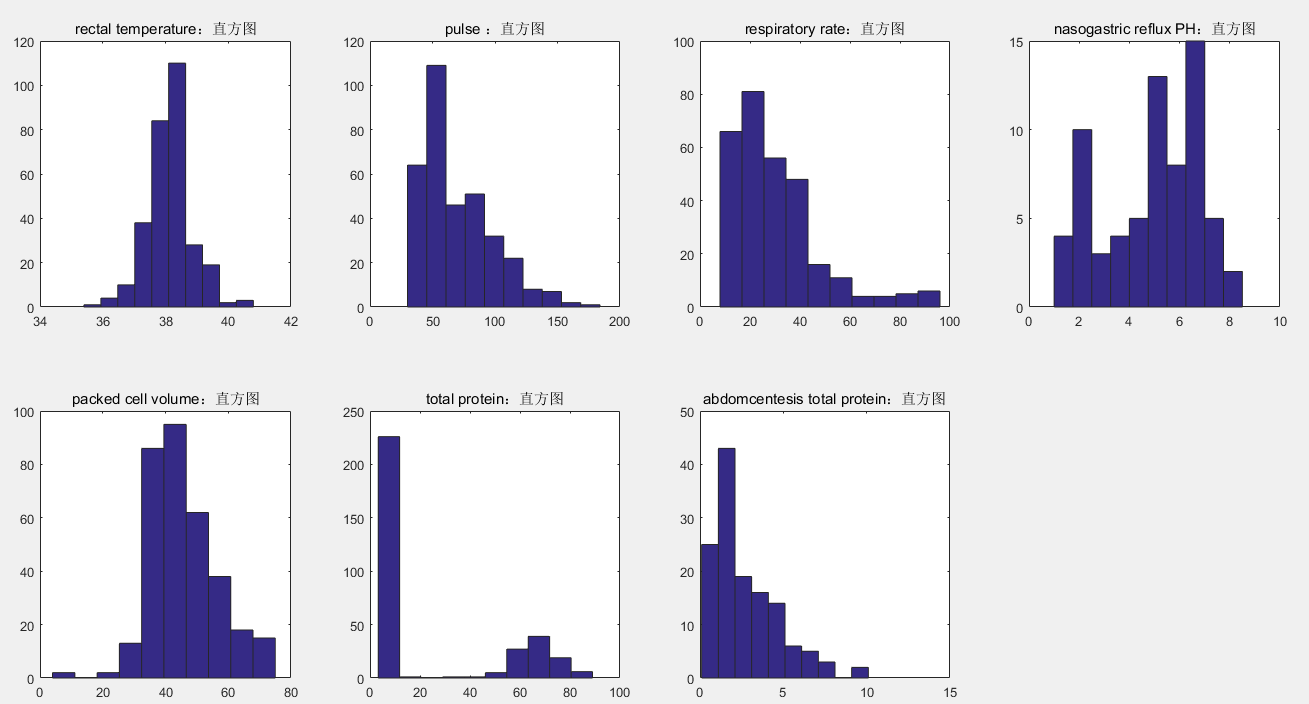
结果如下：



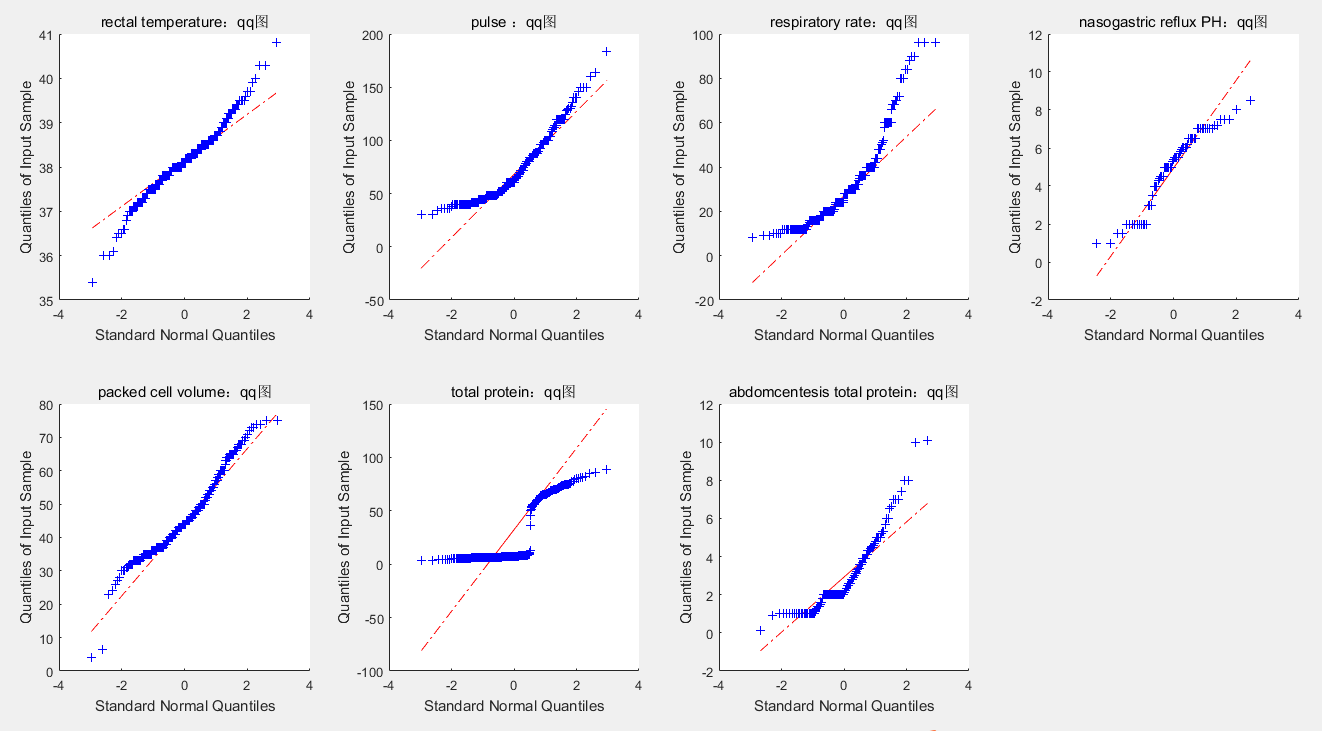
3.绘制直方图和qq图：



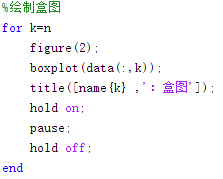
直方图如下：



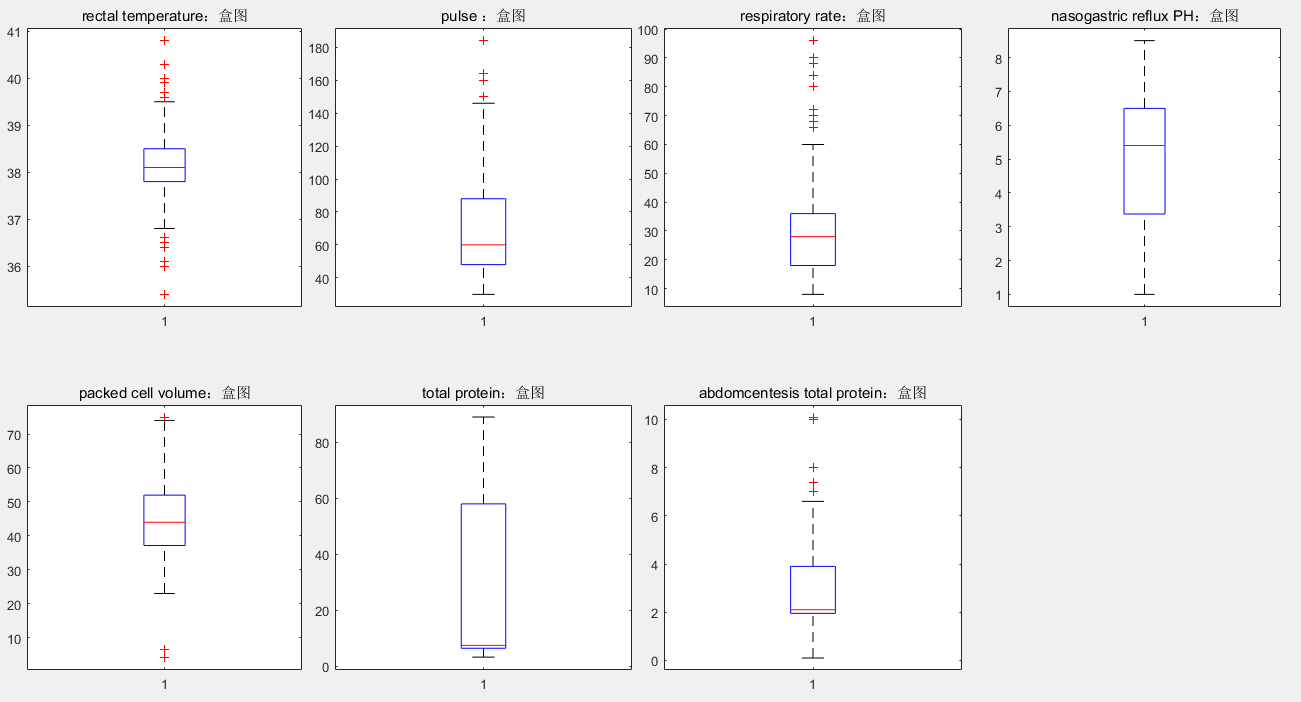
qq图如下：



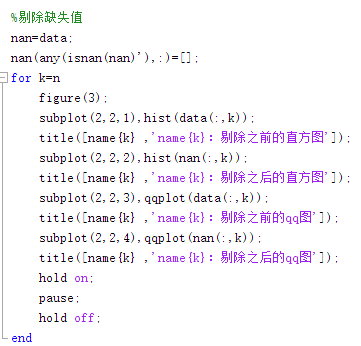
4.绘制盒图：



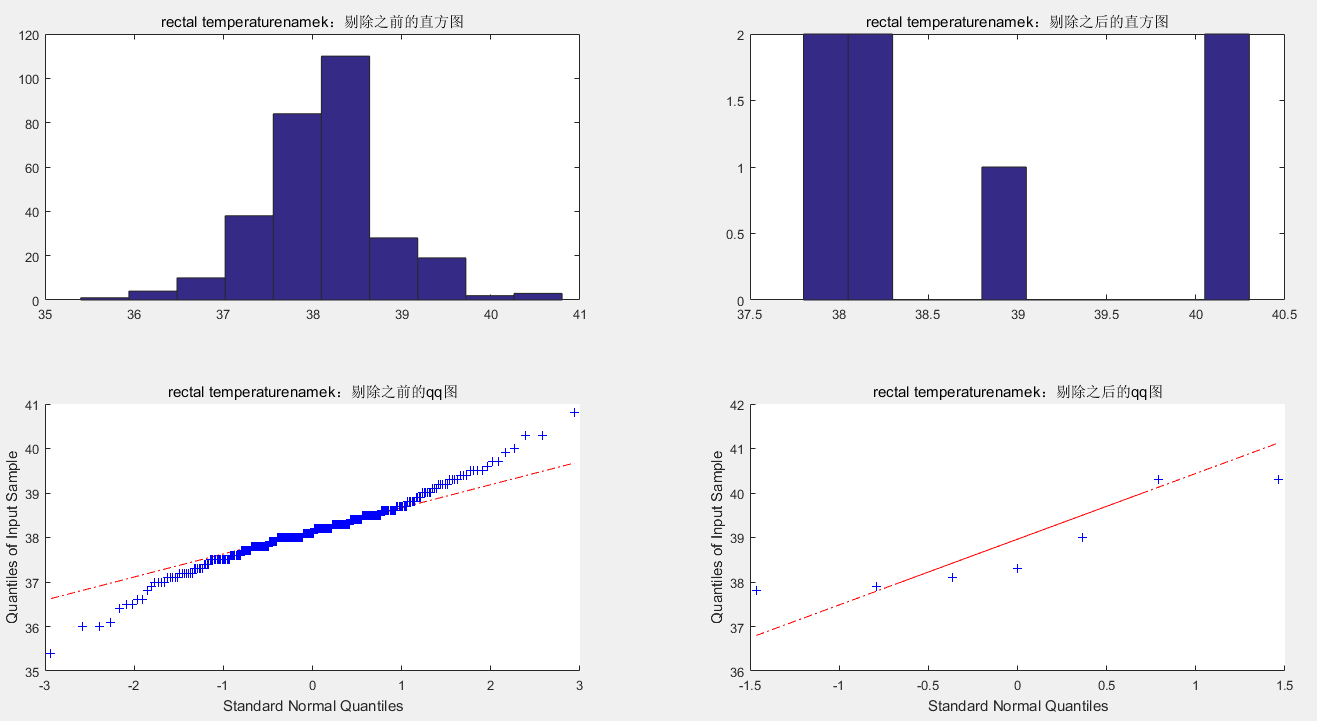
结果如下：

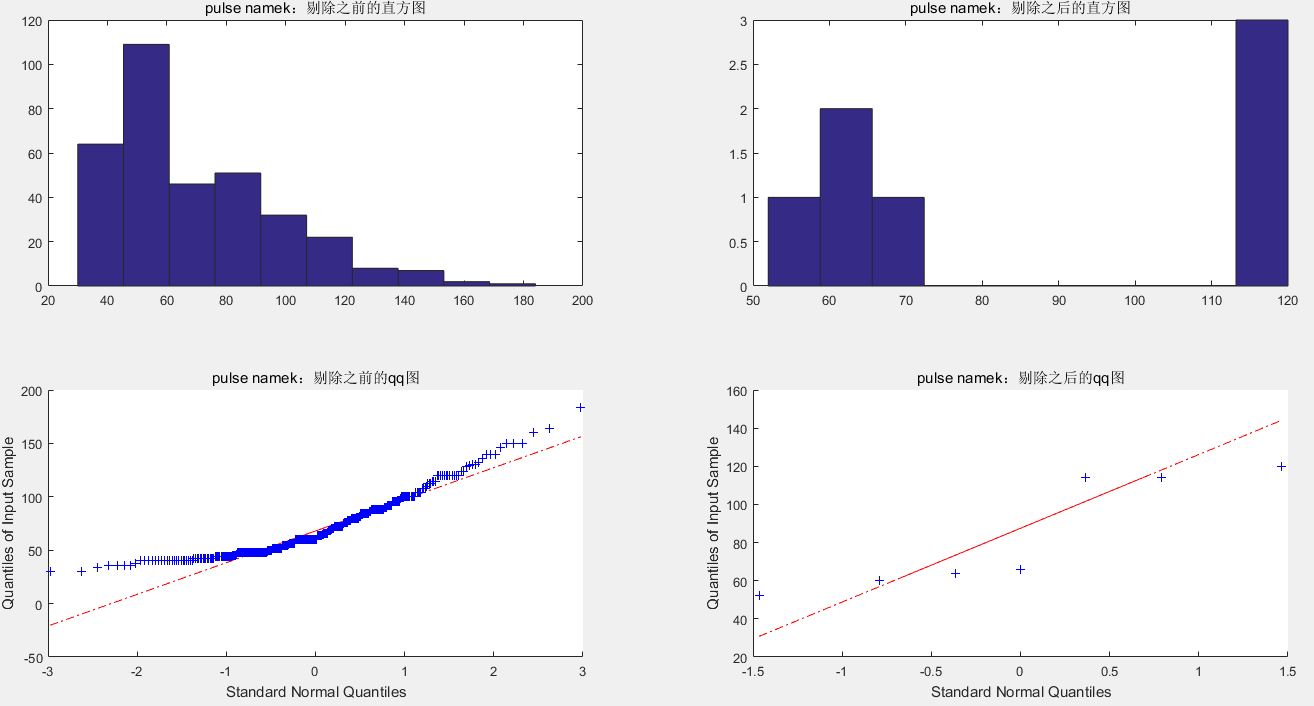


5.将缺失值剔除：



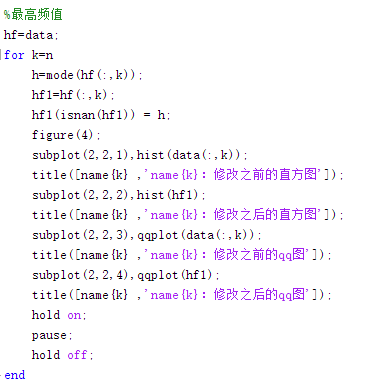
和剔除之前相对比：

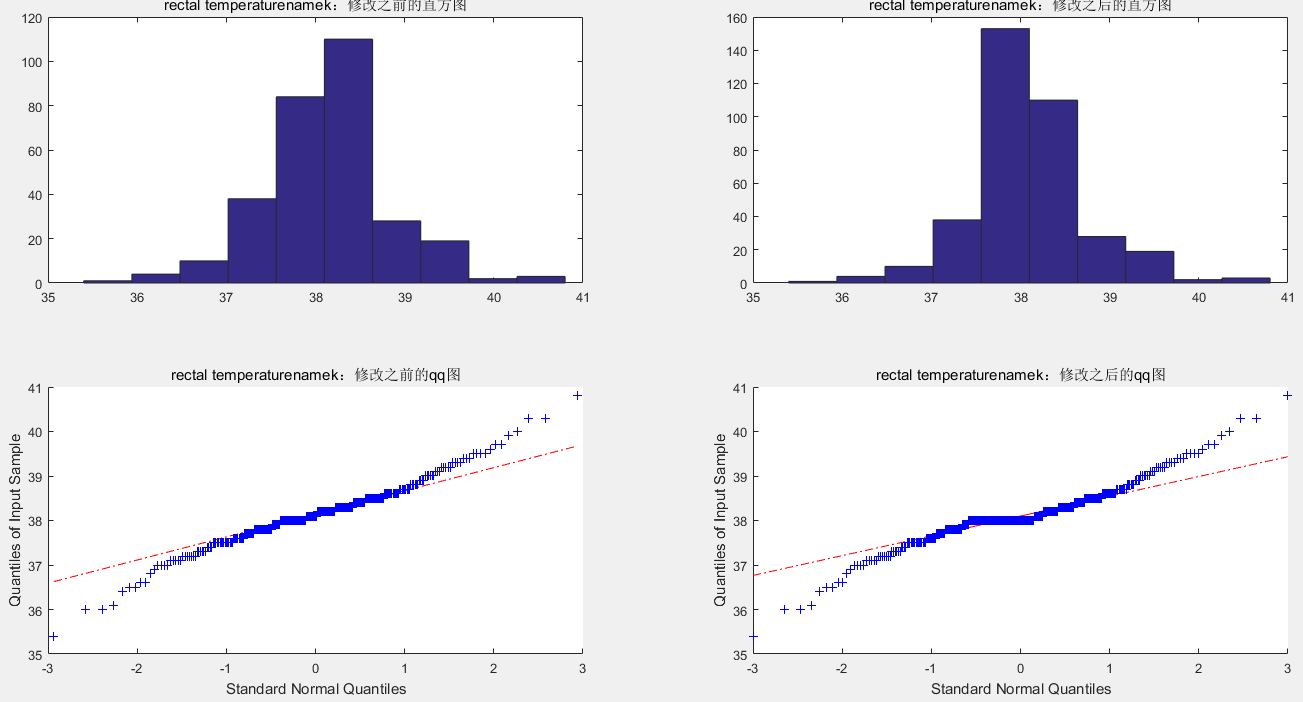


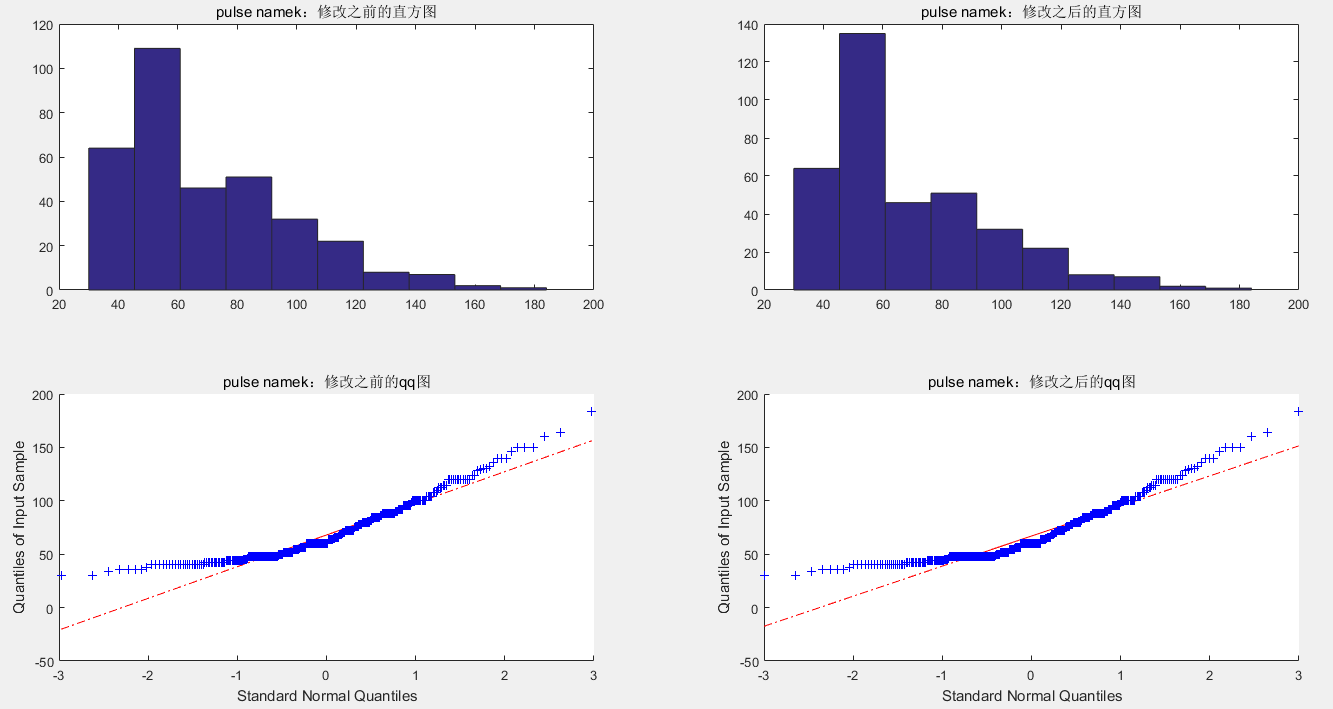


处理之后的数据集保存在nonan.xls中。

6.用高频率来填补缺失值：



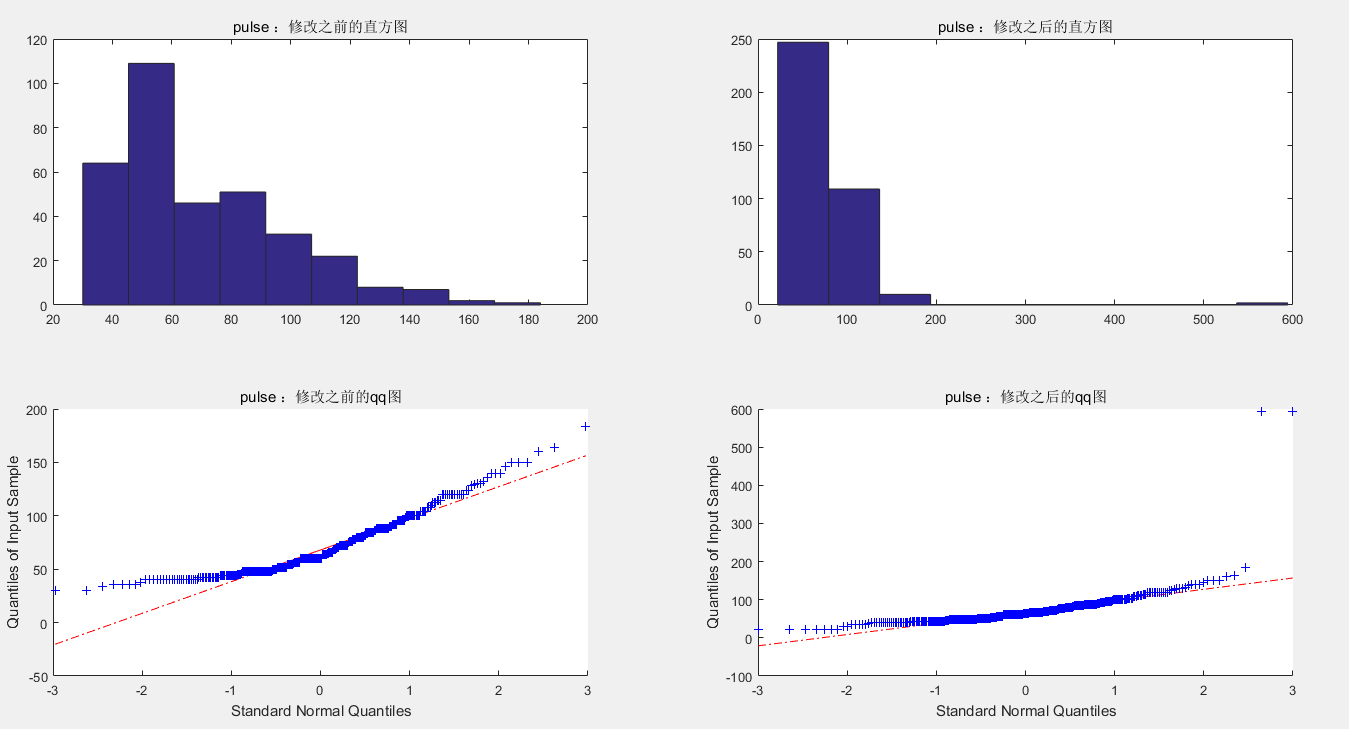


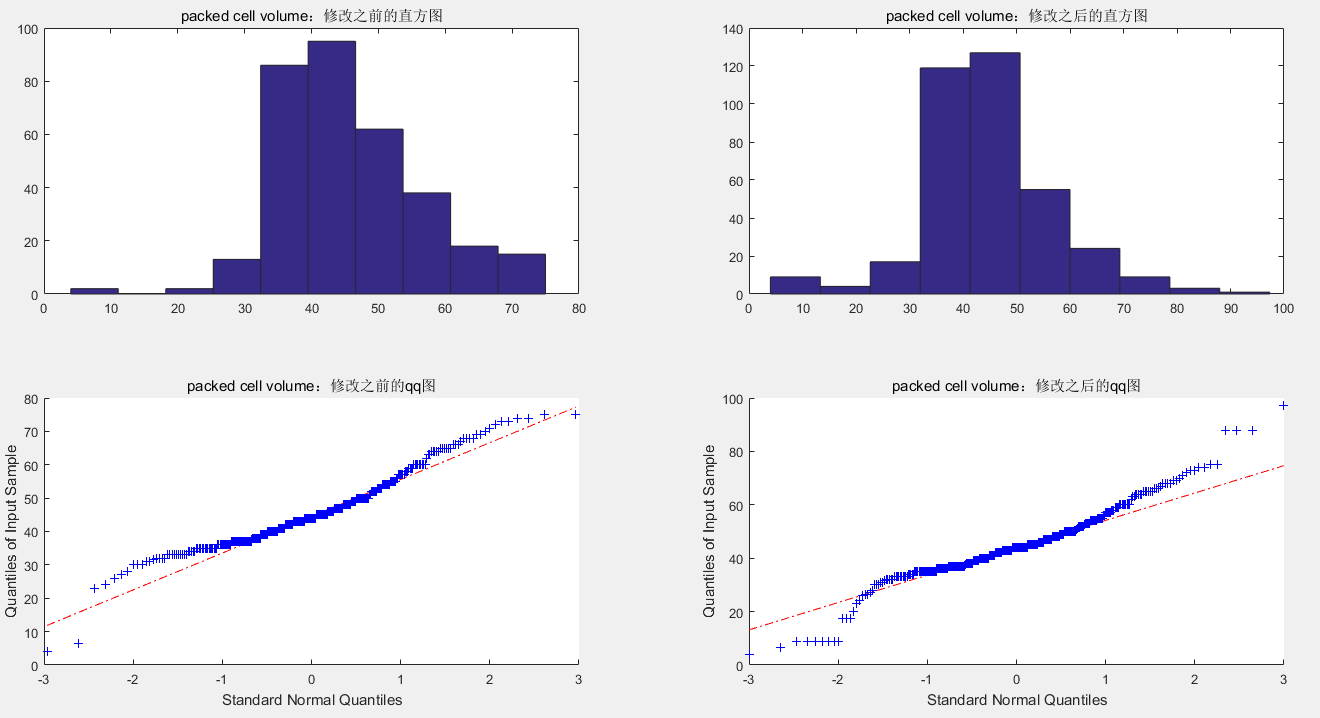


7.通过属性的相关关系填补缺失值：



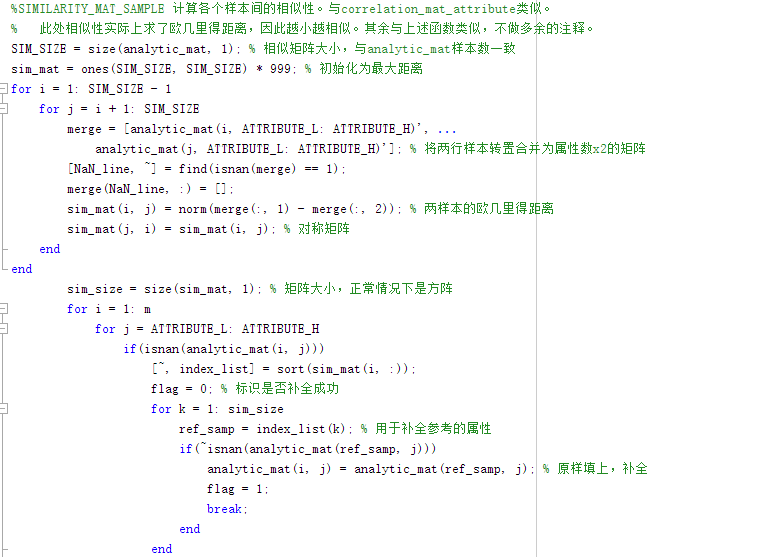
与修改之前的结果相对比：



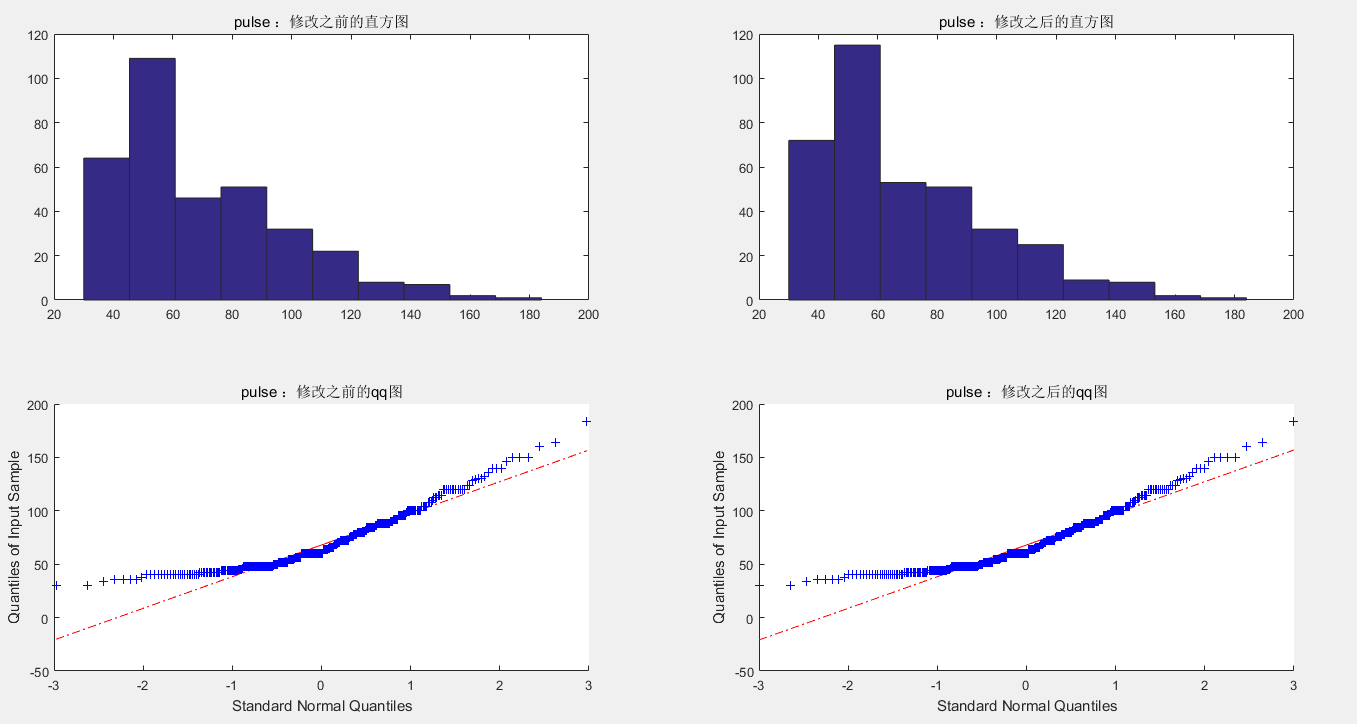


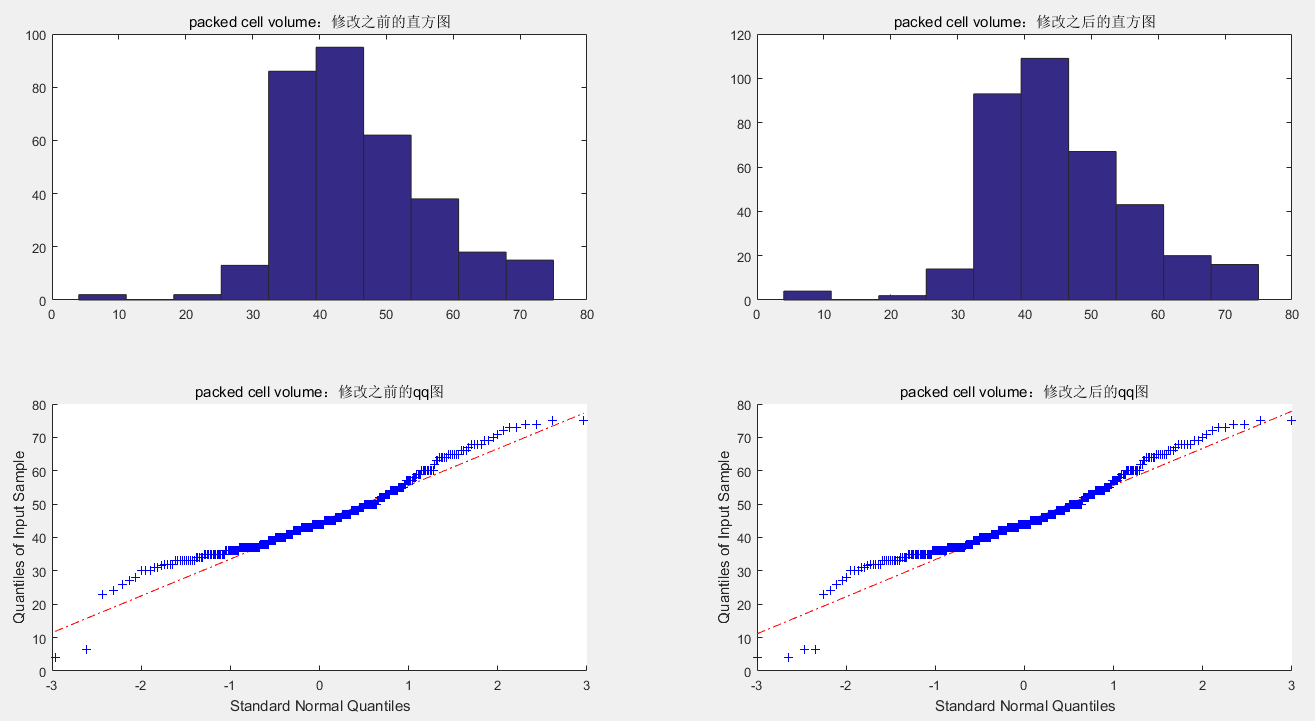
处理之后的数据集保存在cor.xls中。

8.通过案例之间的相似关系填补缺失值



与修改之前的结果先对比：





处理之后的数据集保存在sim.xls中。