## 数据挖掘第一次大作业

## 题目：马的疝病分析

学院：自动化学院

学号：2120160943

姓名：王颖

### 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

共368个样本，27个特征。关于特征的详细说明见下载链接。

### 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。

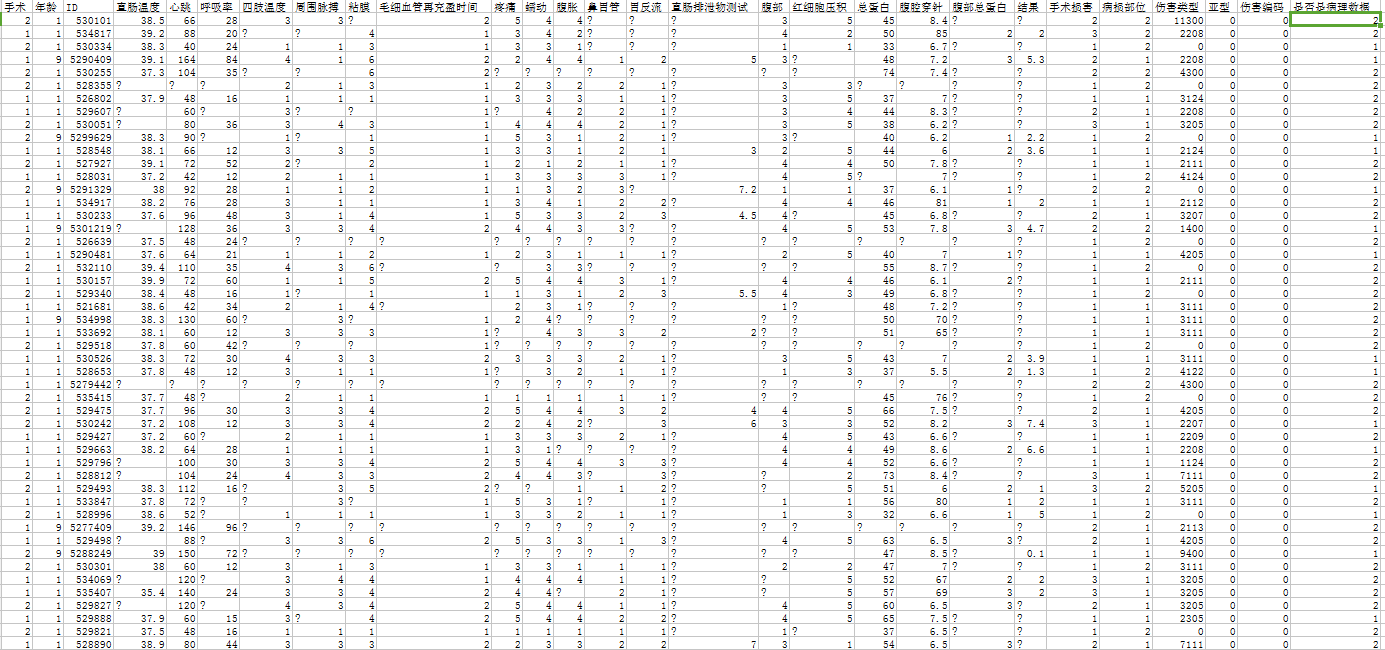
分别使用下列四种策略对缺失值进行处理:

* 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

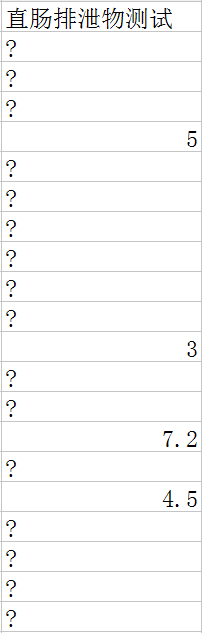
处理后，可视化地对比新旧数据集。

第一部分：

1. 数据的Excel处理和初步分析



首先对数据进行Excel表格化处理，但是通过对数据Excel处理之后发现数据是存在问题的，比如对于第17个属性：直肠排泄物测试：（可能的值有：1=正常 2=上涨 3=下降 4=缺失（缺失的可能代表了一种阻碍）），但是实际呈现的数据里面却是线性连续值，如图：



类似的如下属性全都有类似的问题：

17.直肠排泄物测试：

可能的值有：1=正常 2=上涨 3=下降 4=缺失（缺失的可能代表了一种阻碍）

18.腹部：

可能的值有：1=正常 2=其他 3=顽固的排泄物在大的内脏里面 4=膨胀的小的内脏 5=大内脏膨胀 （3可能是药物治疗带来的阻碍，可以通过药物治愈，4和5代表着一个手术的伤害）

19.红细胞压积：（线性、正常值是30到50，当循环减弱或者脱水的时候会上涨）

20.总蛋白：（线性、正常值的范围是6-7.5（gms/dL）的范围，值越高代表脱水越严重）

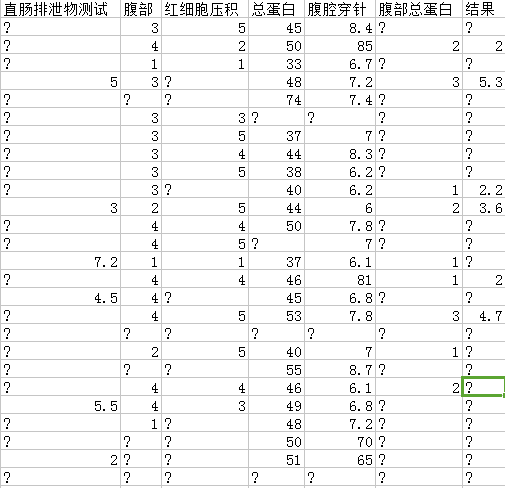
21.腹腔穿刺术的表面：（针穿进妈的腹部以获取来自腹腔的液体）

可能的值有：1=清澈 2=浑浊的 3=血清血液的(清澈的才代表正常，后两个代表蠕动变弱）

22.腹部总蛋白：（线性、越高的值代表着越弱的内脏）

23.结果：（马的最终结果是什么）

可能的值有： 1=或者 2=死了 3=使用了安乐死



上面对应的属性和表格中呈现的数据范围值严重不符！所以相比之下，放弃对于错误数据的处理仅对目前来看没有错误的数据进行处理！

1. 对标称性数据进行可能数据的频数统计

除了第三个ID号码并且去除错误属性数据以外，标称型属性一共有12项。

首先将Excel文件转化为CSV文件：

第一个属性是：手术（二值） 为1 的有180个，为2的有119，缺失值为1

第二个属性为：年龄（二值） 为1的有276个，为9的有24

第一个属性和第二个属性没有缺失值，但是之后的属性都是有缺失值的

第七个属性为：四肢温度（非二值）可能的值有:1=正常 2=温暖 3=凉 4=冷

为1的有78，为2的有30，为3的有109，为4的有27，缺失值有56个

第八个属性：周围脉搏（非二值）可能值有：1=正常 2=上涨 3=下降 4=不足

为1的有115，为2的有5，为3的有103，为4的有8，缺失值有69

第九个属性：粘膜（非二值） 可能的值有1=正常的粉色 2=亮粉色 3=暗淡的粉色

4=苍白的青紫 5=亮红色（感染的） 6=黑青紫。

为1的有79，为2的有30，为3的有58，为4的有41，为5的有25，为6的有10，缺失值有47

第十个属性：毛细血管再充盈时间（二值）

为1的有188，为2的有78，缺失值为32

第十一个属性：疼痛程度（非二值）可能的值有：1=没有疼痛 2=沮丧 3=间歇性轻微疼痛 4=间歇性严重疼痛 5=持续的严重疼痛

分别为：38 59 67 39 42 缺失值为55

第十二个属性：蠕动（非二值）可能的值：1=超级能动 2=正常 3=海波能动的 4=缺少

分别为：39 16 128 73 缺失值为44

第十三个属性：腹胀（非二值）可能的值为：1=无 2=轻微 3=中等的 4=严重的

分别为：76 65 65 38 缺失值为100

第十四个属性：鼻胃管（非二值）可能的值有：1=无 2=轻微 3=多

分别为： 71 102 23 缺失值为：104

第十五个属性：胃反流（非二值）1=无 2=大于一公升 3=小于一公升

分别为： 120 35 39 缺失值为：106

第二十八个属性：是否是病理数据（二值）

分别为： 99 201 缺失值为：0

1. 数值属性，给出最大、最小、均值、中位数及缺失值的个数

有效数据的数值属性有三个：

第四个属性：直肠温度

最大值为40.8 最小值为35.4 均值为：38.17 中位数为：38.2

缺失值有60个

第五个属性为：心跳

最大值为：184.0 最小值为：30.0 均值为：71.91 中位数为：64.0

缺失值个数为：24

第六个属性为：呼吸率

最大值为： 96 最小值为：8 均值为：30.4 中位数为：24.5

缺失值个数为：58

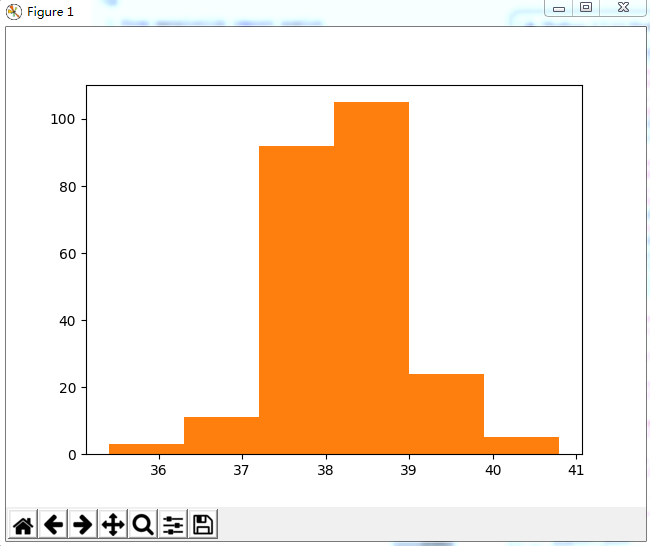
1. 针对数值属性：

1>绘制直方图，如mxPH，用qq图检验其分布是否为正态分布

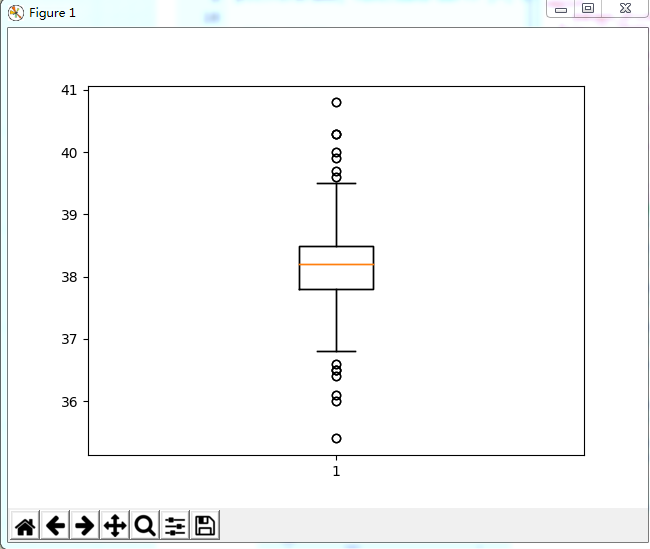
2>绘制盒图，对离群值进行识别

对于第一个属性：

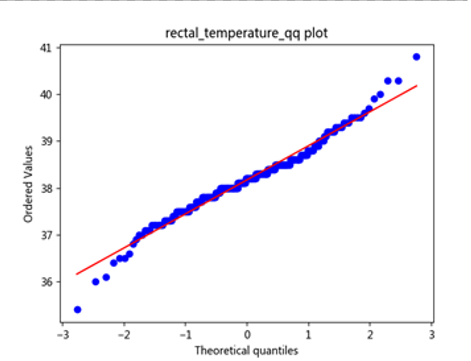
直方图：



盒图：

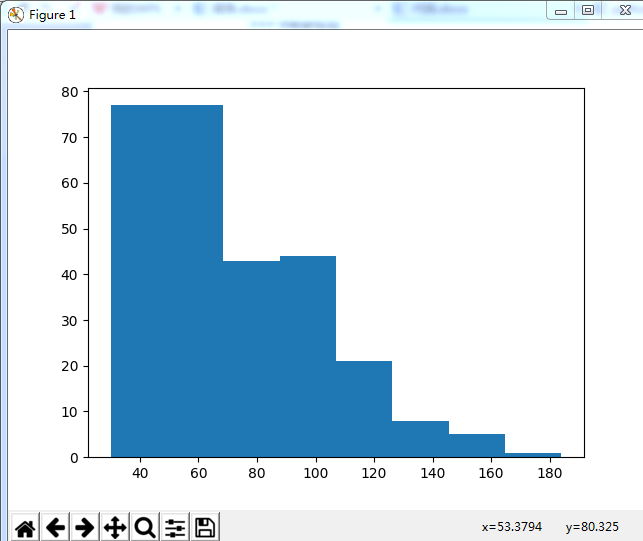


qq图：

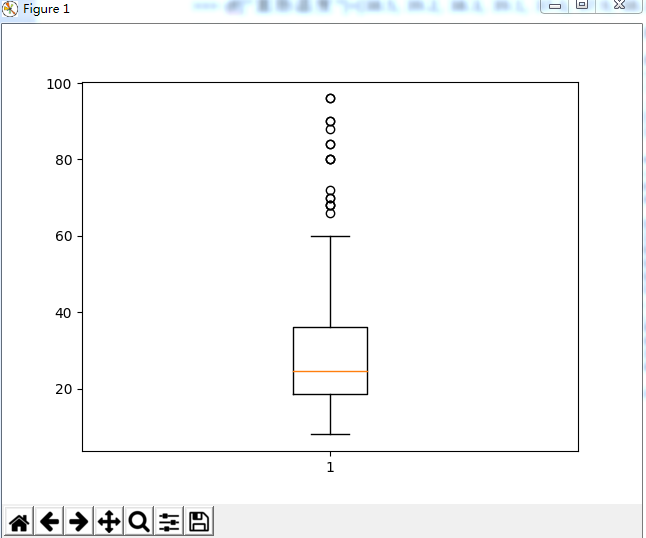


第二个属性：

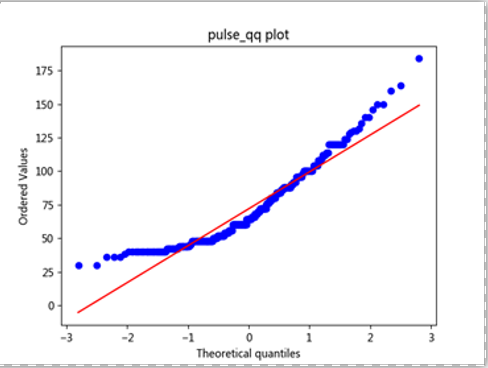
直方图：



盒图：

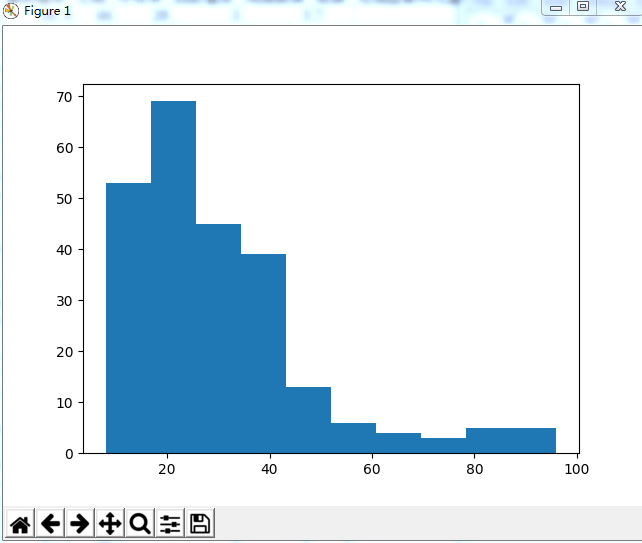


qq图：

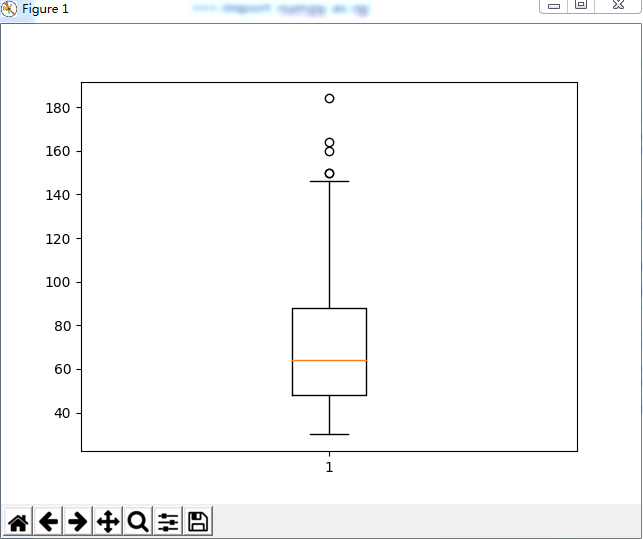


第三个属性：

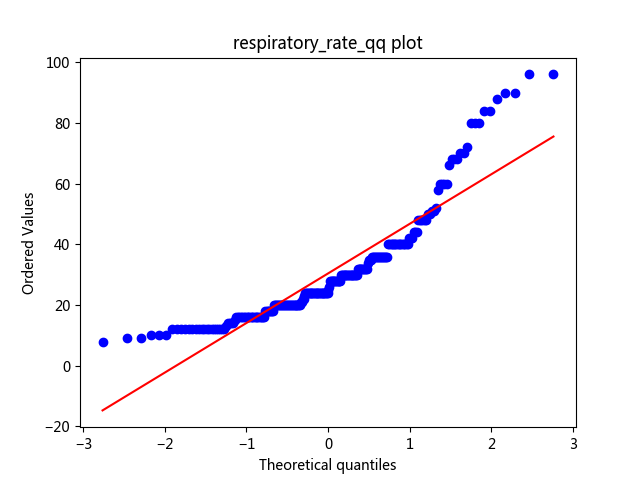
直方图：



盒图：



qq图：



1. 对缺失值进行处理

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

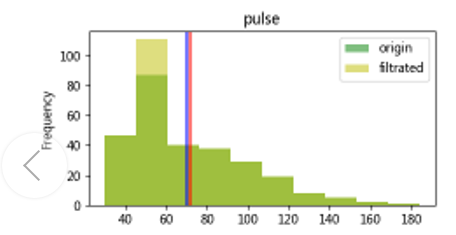
处理后，可视化地对比新旧数据集。

第一：将缺失值部分剔除：

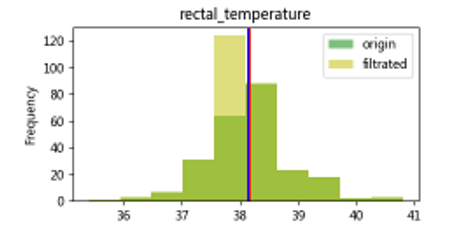
如果采用将所有包含缺失值的行直接进行整行剔除的话，因为包含的数据中所丢失的占的比重比较大，如果整行剔除的话会造成数据损失严重。所以第一种处理办法对本次数据的处理并不合适。

第二：最高频率值来填补缺失值

此方法适用于标称值属性属性（共12个），而通过对数据本身进行观察，可以发现有两个属性没有缺失值，所以要处理的属性个数一共9个。（取其中1个属性作为样例展示）



第三：通过属性的相关关系来填补缺失值，调用的函数。该方法可以用来处理数值型数据（使用一个属性进行展示）



第四：通过数据对象之间的相似性来填补缺失值（可以处理数据型也可以处理标称性数据）

