**马的疝病分析报告**

**姜林 2120161001**

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368个样本，每个样本具有28个特征，包括：21个标称型变量和7个数值型变量。

1. **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。

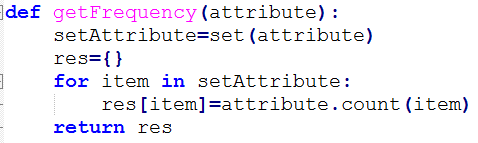
分别使用下列四种策略对缺失值进行处理:

* 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。

1. **数据分析过程**
2. **数据摘要**

对标称属性，给出每个可能取值的频数：

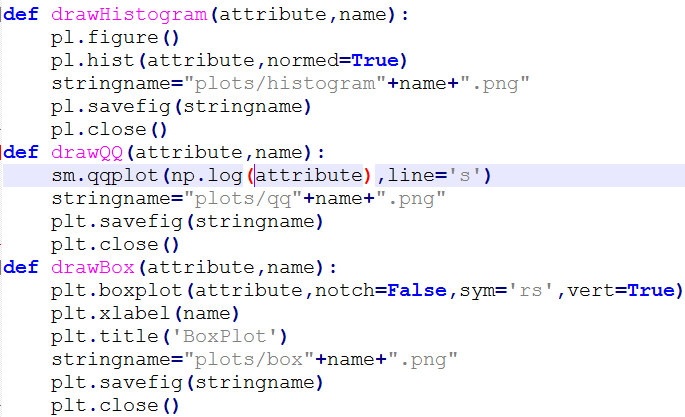


部分结果：

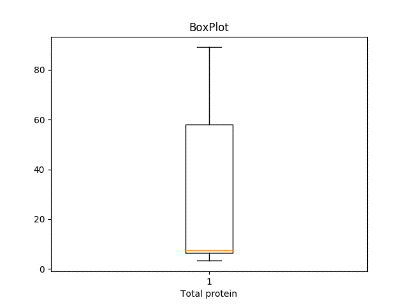
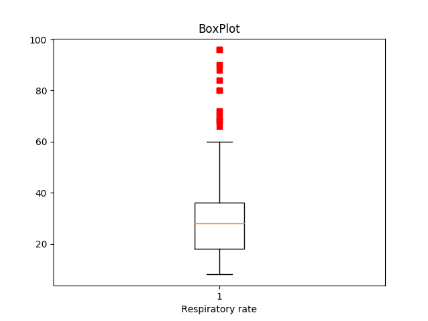
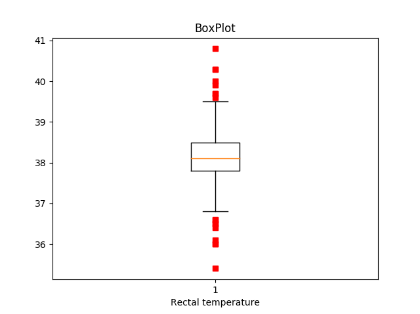
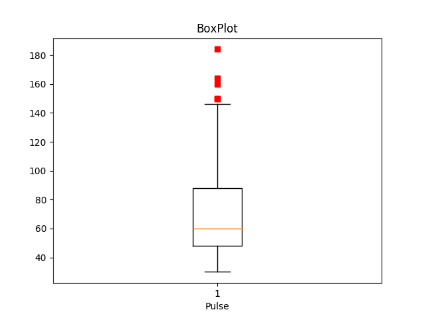
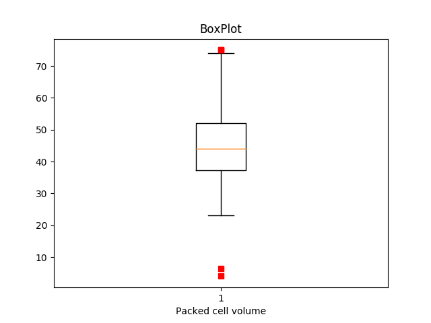
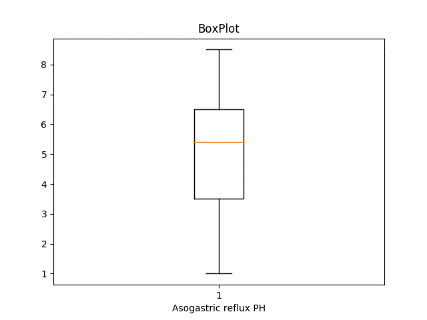
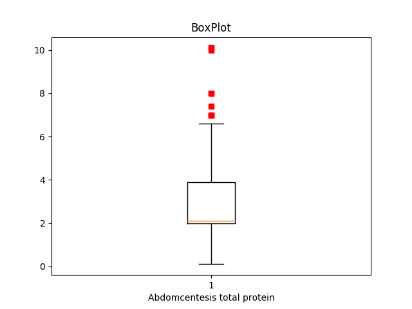
|  |
| --- |
| Surgery The frequency of each value: {'1': 214, '2': 152} |
| Age The frequency of each value: {'1': 340, '9': 28} |
| Peripheral pulse The frequency of each value: {'1': 151, '3': 116, '2': 6, '4': 12} |
| Mucous membranes The frequency of each value: {'1': 98, '3': 81, '2': 38, '5': 28,  '4': 50, '6': 25}  对数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数：    部分结果：  Pulse Max value: 184.0 Min value: 30.0 Average value: 70.7573099415 Median,  Q1,Q3: [60.0, 48.0, 88.0] The number of Miss value: 26  Respiratory rate Max value: 96.0 Min value: 8.0 Average value: 30.5218855219  Median,Q1,Q3: [28.0, 18.0, 36.0] The number of Miss value: 71 |

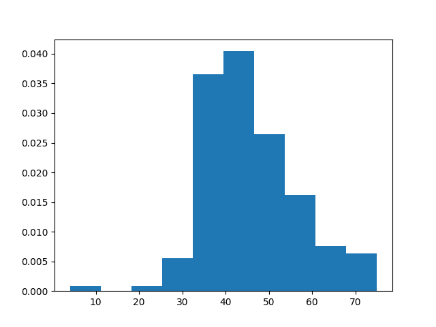
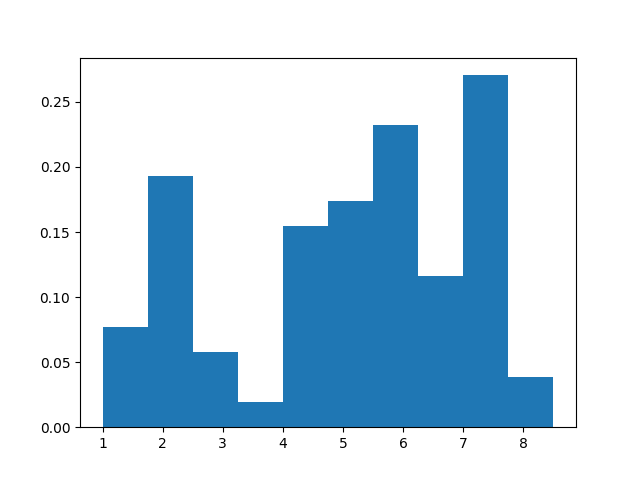
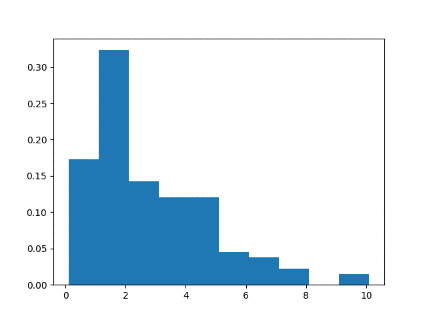
1. **数据可视化**

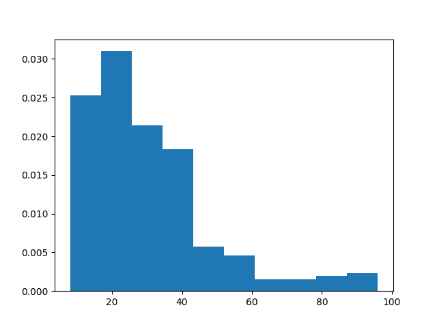
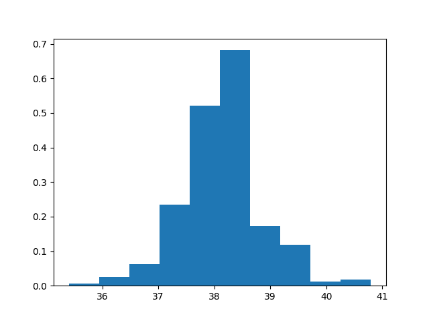
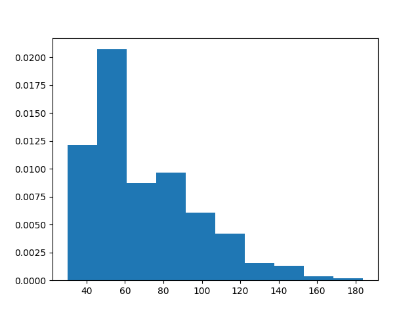
针对数值属性，绘制直方图、qq图、盒图：

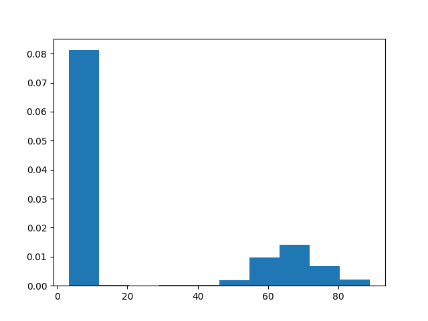


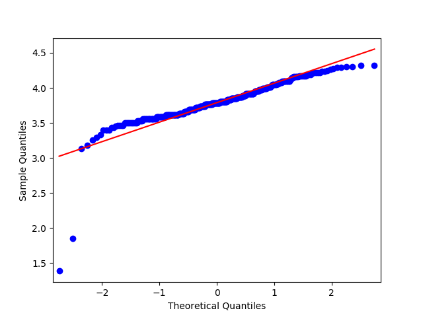
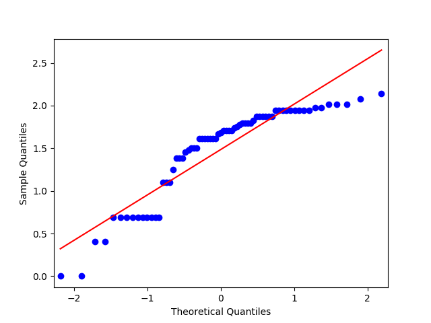
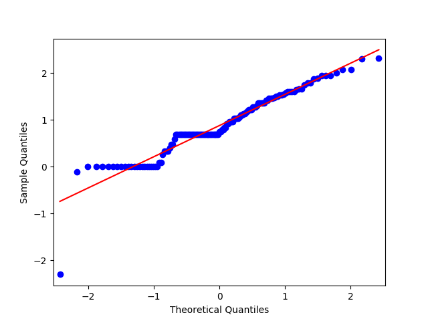
部分结果：

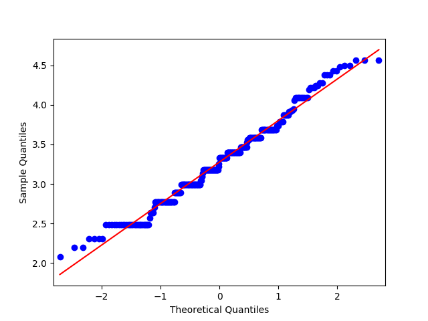
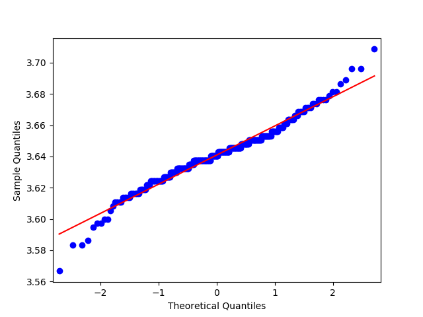
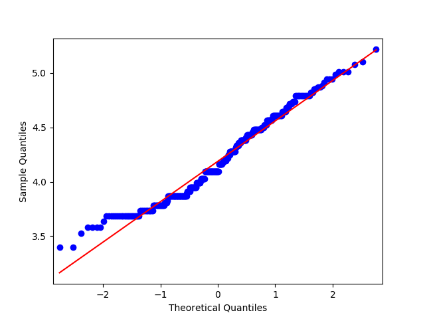


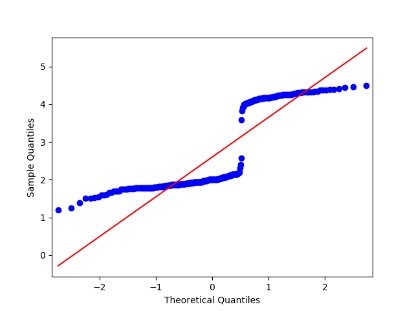












1. **数据缺失处理**