数据挖掘作业 **1** 数据探索性分析与预处理

马的疝病分析

姓名：孙宇超

学号：2120161044

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**

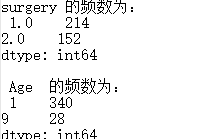
共368个样本，27个特征。

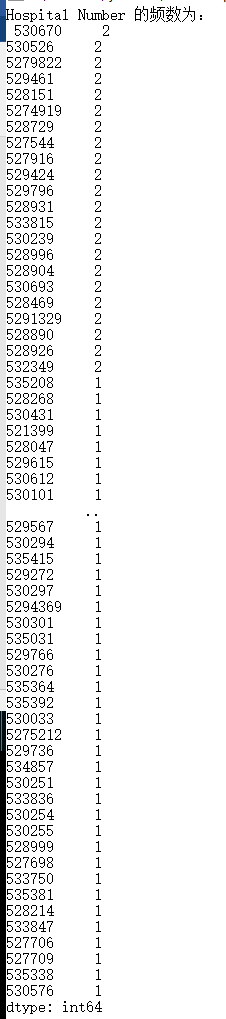
**3. 数据分析要求**

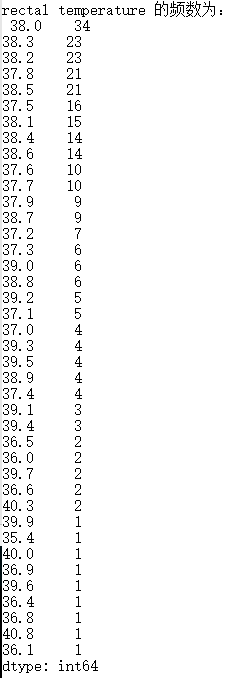
**3.1 数据可视化和摘要**

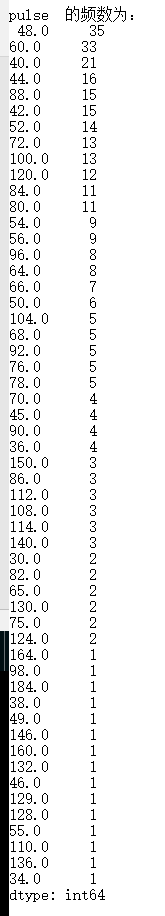
**3.1.1数据摘要**

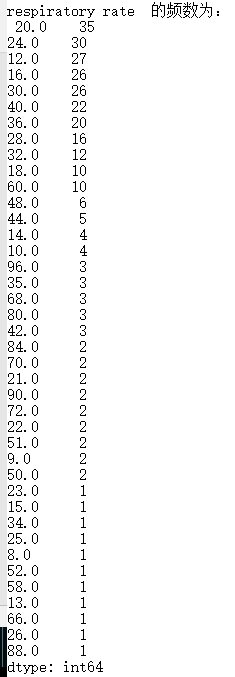
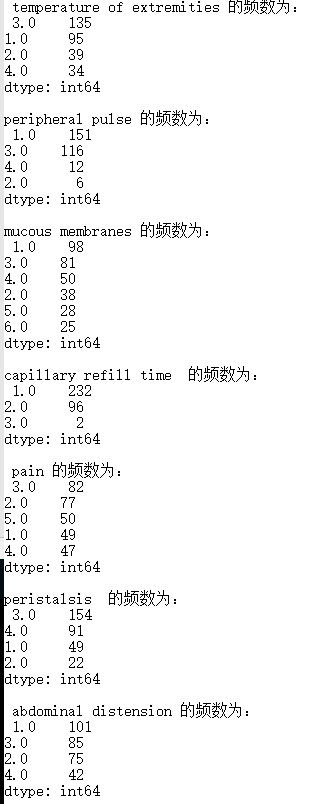
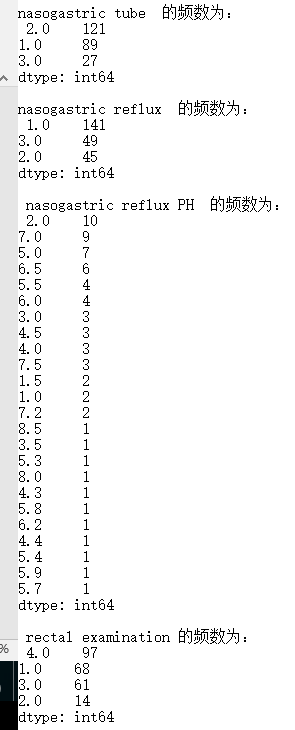
* 对标称属性，给出每个可能取值的频数

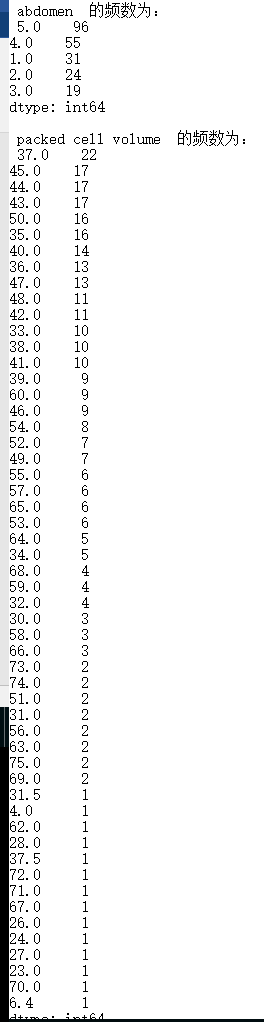


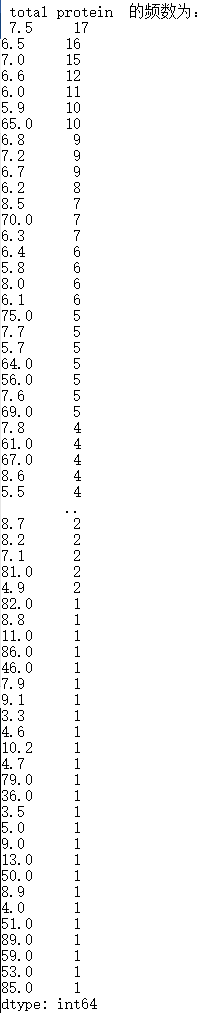


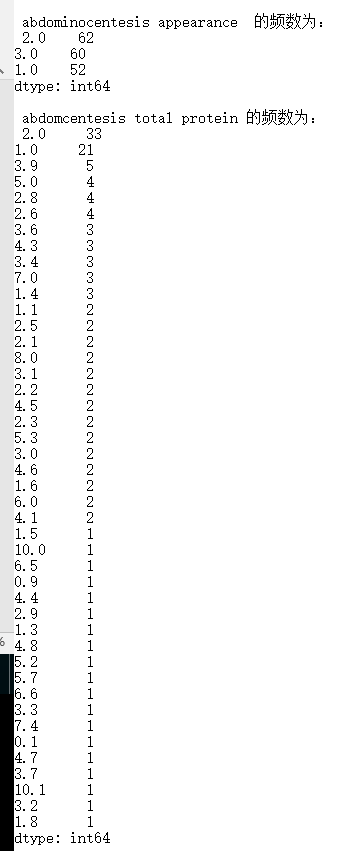


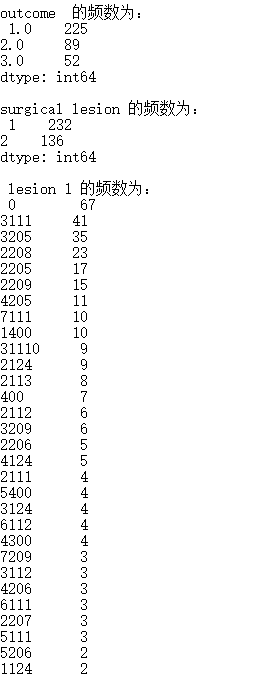




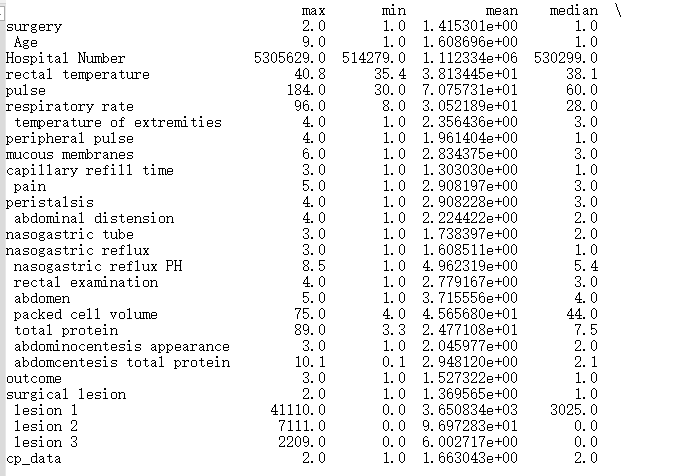
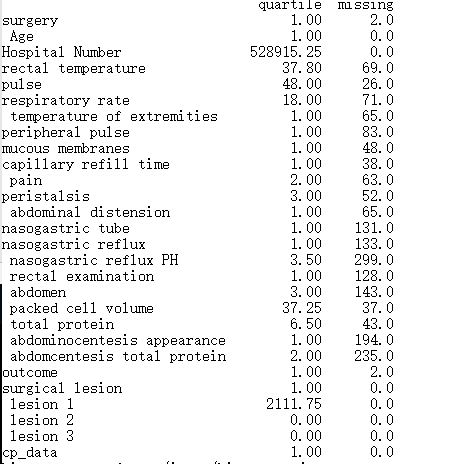








* 数值属性，给出最大、最小、均值、中位数、四分位数及缺失值的个数。

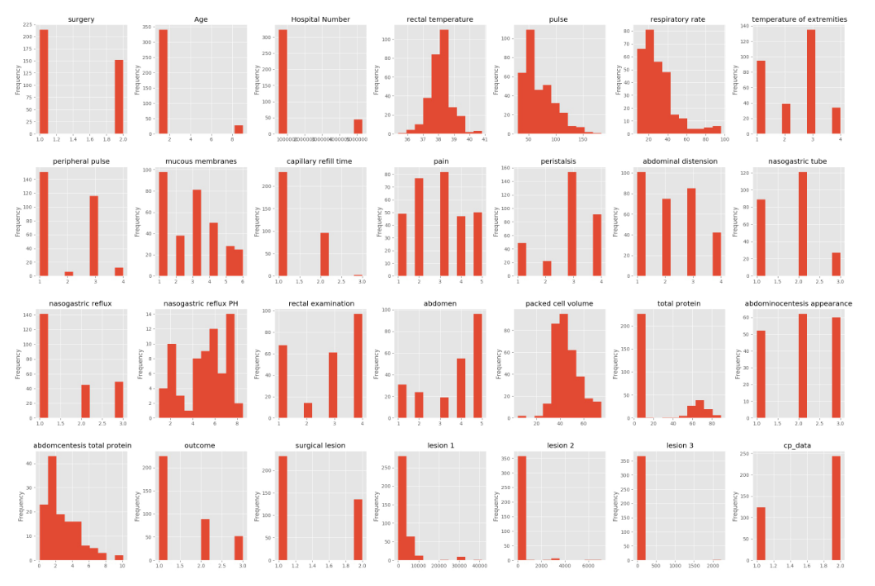


**3.1.2数据的可视化**

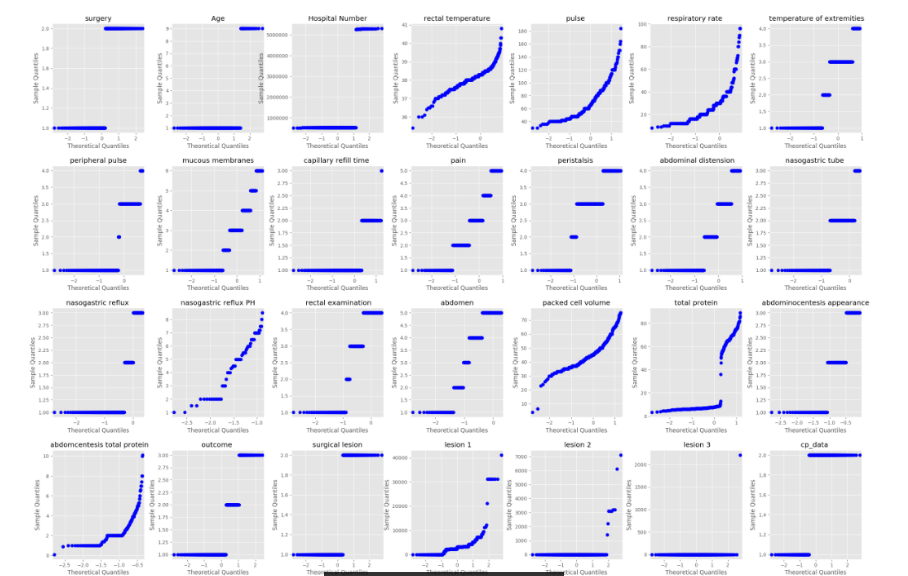
针对数值属性：

* 绘制直方图，如mxPH，用qq图检验其分布是否为正态分布。

直方图如下所示：

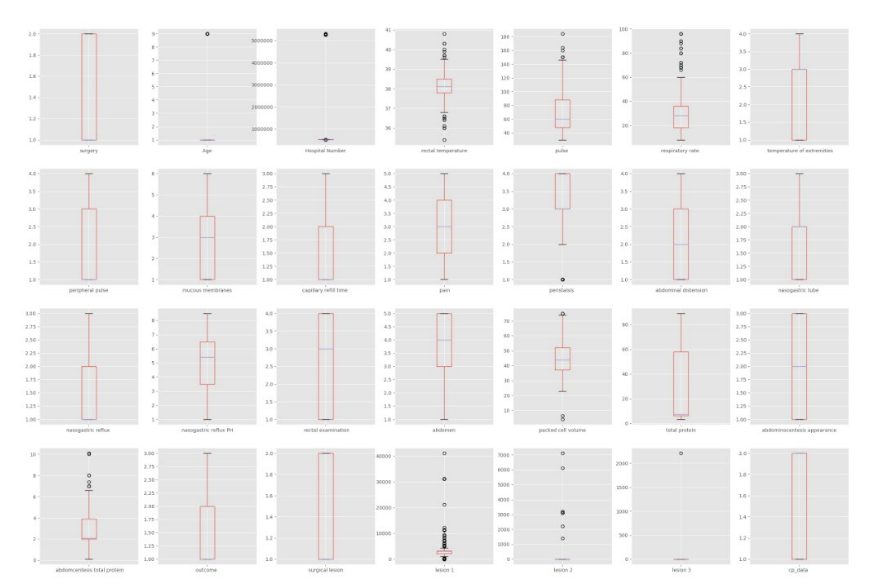


qq图如下所示：



根据各个属性的qq图，得知有两种属性满足正态分布，分别是属性 rectal temperature和属性 packed cell volume。

* 绘制盒图，对离群值进行识别



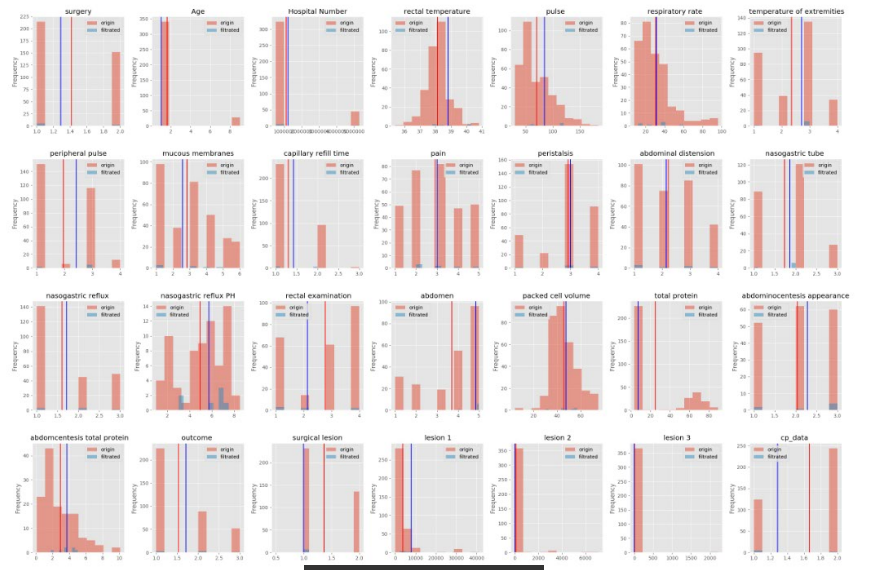
根据各个属性的盒图可以看出，属性Age，Hospital Number，rectal temperature，pulse，respiratory rate，peristalsis，packed cell volume，abdomcentesis total protein，lesion1，lesion2，lesion3存在离群值。

**3.2 数据缺失的处理**

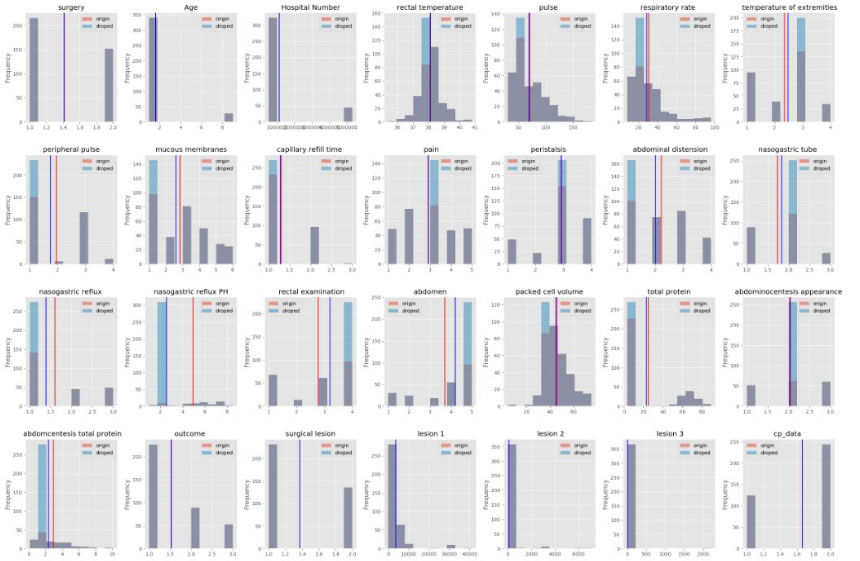
数据集中有30%的值是缺失的，因此需要先处理数据中的缺失值。

分别使用下列四种策略对缺失值进行处理:

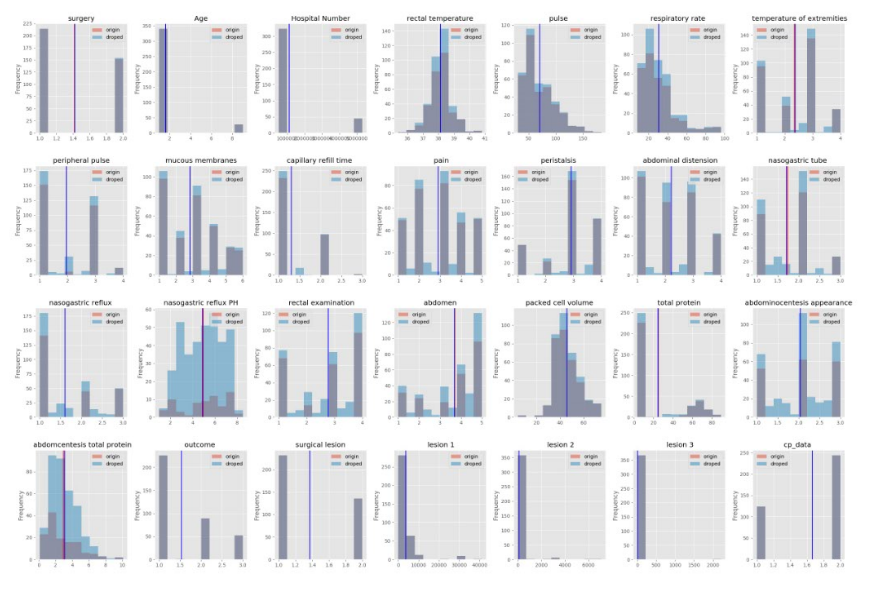
* 将缺失部分剔除



* 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

