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：张力嘉

学号：2120161077

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

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。

4. 提交内容

- 分析过程
- 分析程序

4.1 分析过程

1）数据摘要：将数据从txt转化为excel的csv格式，方便读写。读取csv文件，根据数据集文档进行属性赋值。属性值为："surgery"," Age ","Hospital Number","rectal temperature","pulse ","respiratory rate "," temperature of extremities","peripheral pulse","mucous membranes","capillary refill time "," pain","peristalsis "," abdominal distension","nasogastric tube ","nasogastric reflux "," nasogastric reflux PH "," rectal examination"," abdomen "," packed cell volume "," total protein "," abdominocentesis appearance "," abdomcentesis total protein","outcome ","surgical lesion"," lesion 1"," lesion 2"," lesion 3","cp_data "。

数值属性：" temperature of extremities"，"pulse "，"respiratory rate "，" nasogastric reflux PH "，" packed cell volume "，" total protein "，" abdomcentesis total protein"。
其余为标称属性。

标称属性频数：因为标称属性个数过多，这里只列举了部分数据。

Frequency of surgery attribute:

Value	Count	Percent
1	214	58.47%
2	152	41.53%

Frequency of Age attribute:

Value	Count	Percent
1	340	92.39%
2	0	0.00%
3	0	0.00%
4	0	0.00%
5	0	0.00%
6	0	0.00%
7	0	0.00%
8	0	0.00%
9	28	7.61%

Frequency of temperature of extremities attribute:

Value	Count	Percent
1	95	31.35%
2	39	12.87%
3	135	44.55%
4	34	11.22%

Frequency of peripheral pulse attribute:

Value	Count	Percent
1	151	52.98%
2	6	2.11%
3	116	40.70%
4	12	4.21%

Frequency of mucous membranes attribute:

Value	Count	Percent
1	98	30.63%
2	38	11.88%
3	81	25.31%
4	50	15.63%
5	28	8.75%
6	25	7.81%

Frequency of capillary refill time attribute:

Value	Count	Percent
1	232	70.30%
2	96	29.09%
3	2	0.61%

Frequency of pain attribute:

Value	Count	Percent
1	49	16.07%
2	77	25.25%
3	82	26.89%
4	47	15.41%
5	50	16.39%

Frequency of peristalsis attribute:

Value	Count	Percent
1	49	15.51%
2	22	6.96%
3	154	48.73%
4	91	28.80%

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：

Data abstract of attribute rectal temperature:
Maximium: 40.8
Minimium: 35.4
Average: 38.1344
Median: 38.1
Quartile: 37.8, 38.5
Missing data: 69

Data abstract of attribute peripheral pulse:
Maximium: 184
Minimium: 30
Average: 70.7573
Median: 60
Quartile: 48, 88
Missing data: 26

Data abstract of attribute respiratory rate:
Maximium: 96
Minimium: 8
Average: 30.5219
Median: 28
Quartile: 18, 36
Missing data: 71

Data abstract of attribute nasogastric reflux PH:
Maximium: 8.5
Minimium: 1
Average: 4.9623
Median: 5.4
Quartile: 3.375, 6.5
Missing data: 299

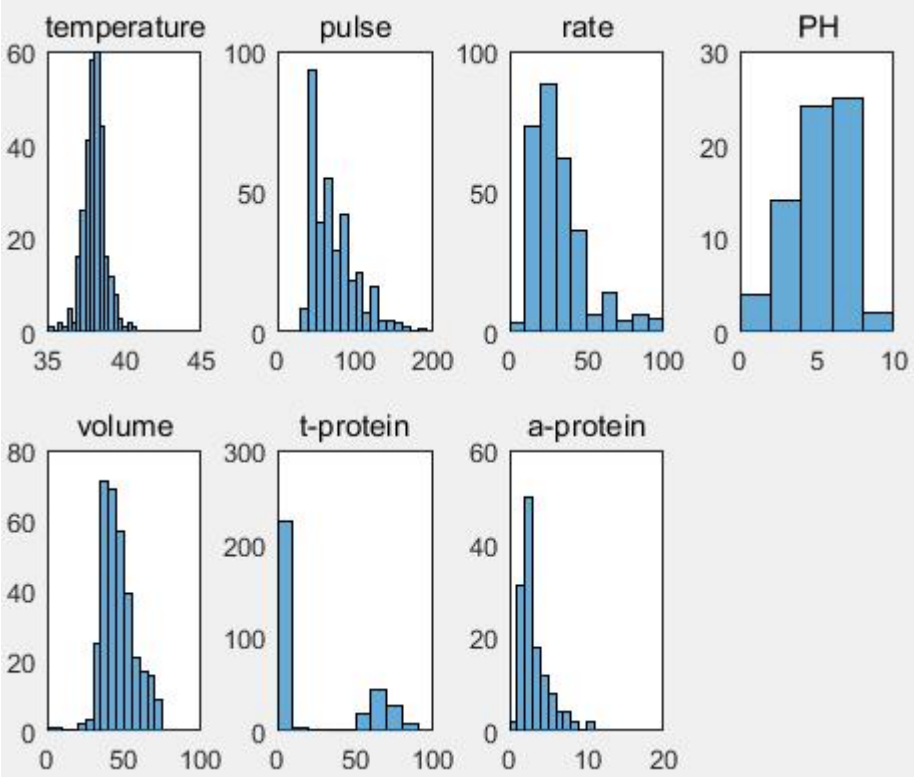
Data abstract of attribute packed cell volume:
Maximium: 75
Minimium: 4
Average: 45.6568
Median: 44
Quartile: 37.125, 52
Missing data: 37

Data abstract of attribute total protein:
Maximium: 89
Minimium: 3.3
Average: 24.7711
Median: 7.5
Quartile: 6.5, 58
Missing data: 43

Data abstract of attribute abdomcentesis total protein:
Maximium: 10.1
Minimium: 0.1
Average: 2.9481
Median: 2.1
Quartile: 1.95, 3.9
Missing data: 235

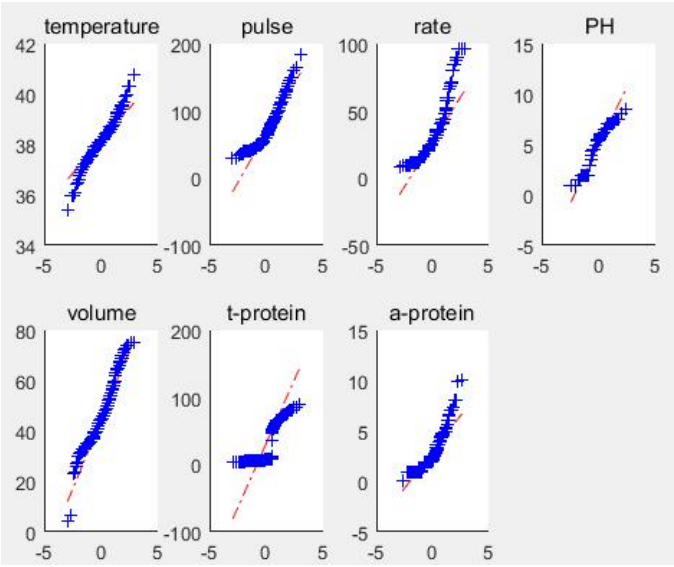
2）数据可视化

数值属性直方图：

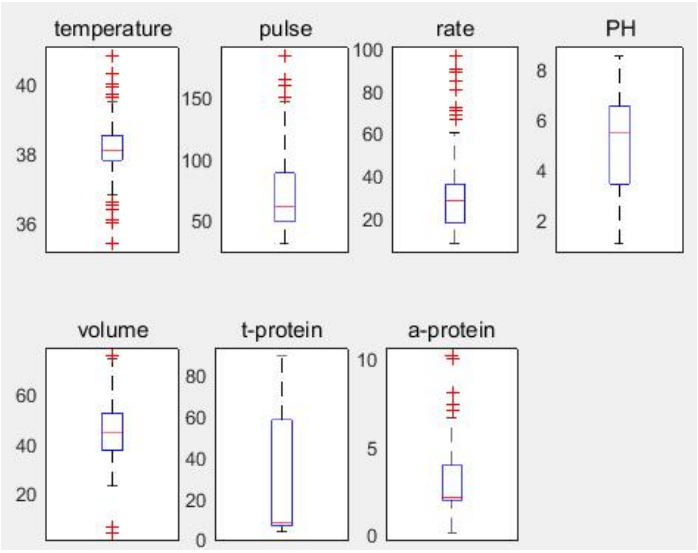


由图可看出除了t-protein的分布之外其他分布都接近正态分布，由QQ图更精准的判定其他几个性质分布和正态分布的相似度。

QQ图：由下面的QQ图可以更加精确的看出temperature和volume两个属性更加接近正态分布。



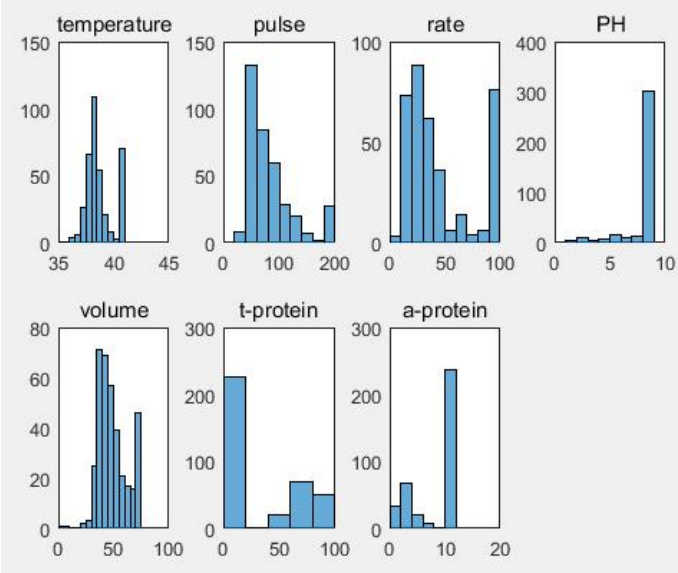
盒图：由以下盒图可以看出temperature属性和rate属性相比其他属性具有较多的离群值。



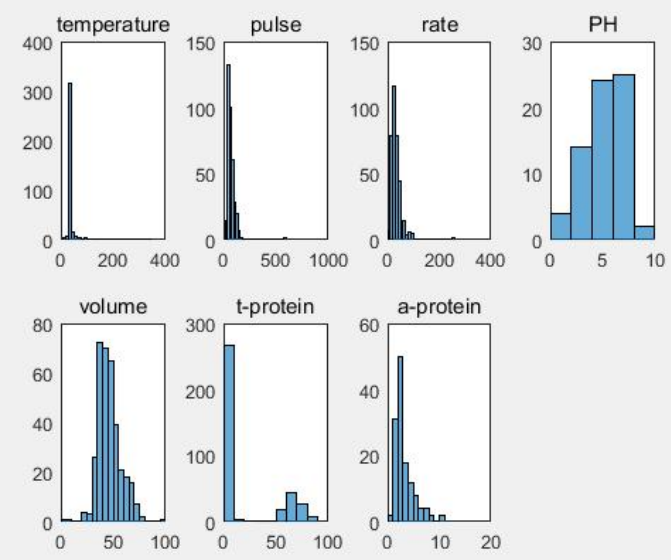
3）数据集预处理

1.剔除缺省值的操作已经在之前做好，在此不再赘述。

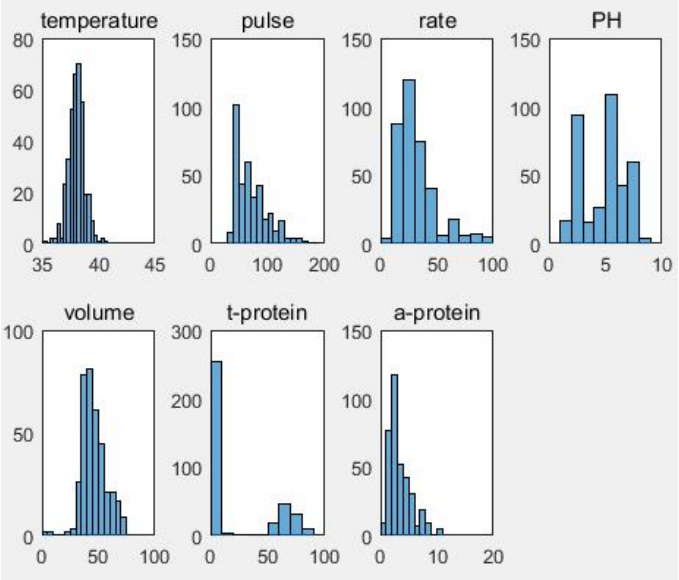
2.用最高频率值来填补缺失值：在此属性中的缺失值用此属性中所计算出的最高频率值填补。直方图如下



3.通过属性的相关关系来填补缺失值：计算两个属性的相关性，相关性越大越可以根据另一个属性推断缺失属性的值。通过另一属性的回归分析，计算当前的缺失值。直方图如下：



4.通过数据对象之间的相似性来填补缺失值：计算两个样本的相似程度，越相似证明越可以使用该样本推断当前含缺失。直方图如下：



由以上四幅图比较可知，按相似性填补的结果和处理前更相似，按用最高频率值来填补缺失值的结果和处理前差别最大。