数据挖掘——马的疝病分析

2120161061 王雨佳

一、数据摘要

为方便处理，首先将数据中缺失数据“?”统一替换为“NaN”，且 Hospital Number 不具有意义，在处理过程中不考虑该属性。

1. 对标称属性，给出每个可能取值的频数。

针对标称属性（surgery/Age/temperature of extremities/peripheral pulse / mucous membranes/capillary refill time/pain/peristalsis/abdominal distension/ nasogastric tube/nasogastric reflux/rectal examination/abdomen/ abdominocentesis appearance/outcome/surgical lesion/lesion1/lesion2/lesion3/cp_data）每个可能取值的频数，利用 matlab 函数“tabulate”进行计算，由于属性较多，随机选择一部分属性进行结果展示，结果如下：

```
>> tabulate(attribute(:,1))
  Value    Count
      1      214
      2      152
```

```
>> tabulate(attribute(:,2))
  Value    Count
      1      340
      9       28
```

```
>> tabulate(attribute(:,7))
  Value    Count
      1       95
      2       39
      3      135
      4       34
```

```
>> tabulate(attribute(:,8))
  Value    Count
      1      151
      2        6
      3      116
      4       12
```

```
>> tabulate(attribute(:,9))
  Value    Count
      1       98
      2       38
      3       81
      4       50
      5       28
      6       25
```

```
>> tabulate(attribute(:,10))
  Value    Count
      1      232
      2       96
      3        2
```

```
>> tabulate(attribute(:,11))
  Value    Count
      1       49
      2       77
      3       82
      4       47
      5       50
```

```
>> tabulate(attribute(:,12))
  Value    Count
      1       49
      2       22
      3      154
      4       91
```

```
>> tabulate(attribute(:,13))
  Value    Count
      1      101
      2       75
      3       85
      4       42
```

```
>> tabulate(attribute(:,14))
  Value    Count
      1       89
      2      121
      3       27
```

```
>> tabulate(attribute(:,15))
  Value    Count
      1      141
      2       45
      3       49
```

```
>> tabulate(attribute(:,17))
  Value    Count
      1       68
      2       14
      3       61
      4       97
```

```
>> tabulate(attribute(:,18))
  Value    Count
      1       31
      2       24
      3       19
      4       55
      5       96
```

```
>> tabulate(attribute(:,21))
  Value    Count
      1       52
      2       62
      3       60
```

```
>> tabulate(attribute(:,23))
  Value    Count
      1      225
      2       89
      3       52
```

```
>> tabulate(attribute(:,24))
  Value    Count
      1      232
      2      136
```

2. 数值属性（rectal temperature/pulse/respiratory rate/nasogastric reflux PH/ packed cell volume/total protein/abdomcentesis total protein），最大、最小、均值、中位数、四分位数及缺失值的个数如下图所示：

rectal_temperature
max: 40.8
min: 35.4
mean: 38.1344
median: 38.1
quartile-1/4: 37.8
quartile-3/4: 38.5
missing: 69

nasogastric_reflux_PH
max: 8.5
min: 1
mean: 4.9623
median: 5.4
quartile-1/4: 3.375
quartile-3/4: 6.5
missing: 299

abdomcentesis_total_protein
max: 10.1
min: 0.1
mean: 2.9481
median: 2.1
quartile-1/4: 1.95
quartile-3/4: 3.9
missing: 235

pulse
max: 184
min: 30
mean: 70.7573
median: 60
quartile-1/4: 48
quartile-3/4: 88
missing: 26

packed_cell_volume
max: 75
min: 4
mean: 45.6568
median: 44
quartile-1/4: 37.125
quartile-3/4: 52
missing: 37

respiratory_rate
max: 96
min: 8
mean: 30.5219
median: 28
quartile-1/4: 18
quartile-3/4: 36
missing: 71

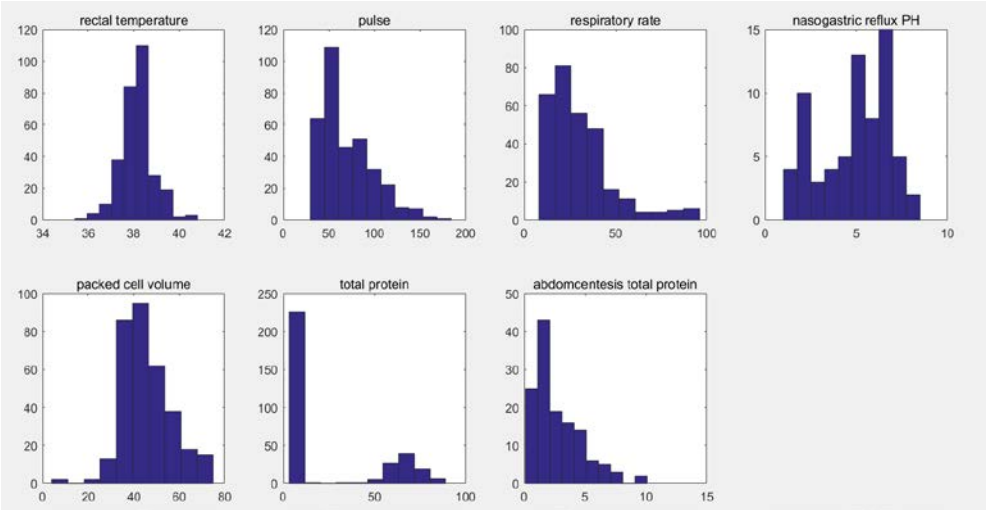
total_protein
max: 89
min: 3.3
mean: 24.7711
median: 7.5
quartile-1/4: 6.5
quartile-3/4: 58
missing: 43

二、数据可视化

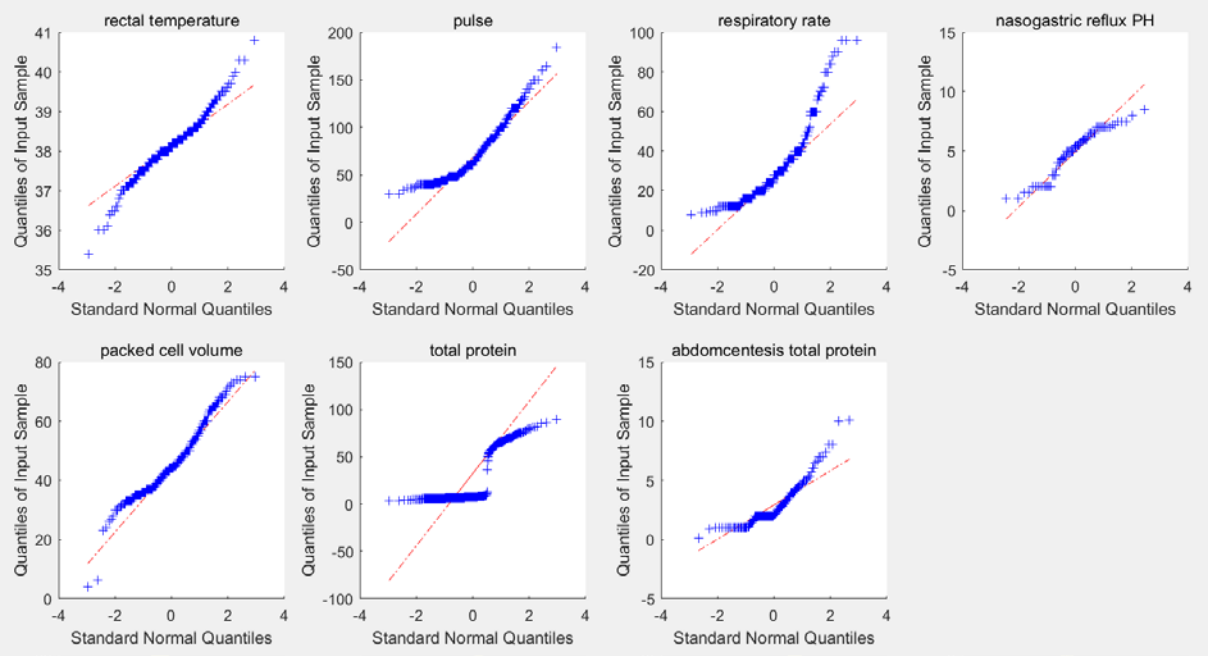
针对数值属性，

1. 绘制直方图，如 mxPH，用 qq 图检验其分布是否为正态分布。

针对各数值属性，直方图如下所示：

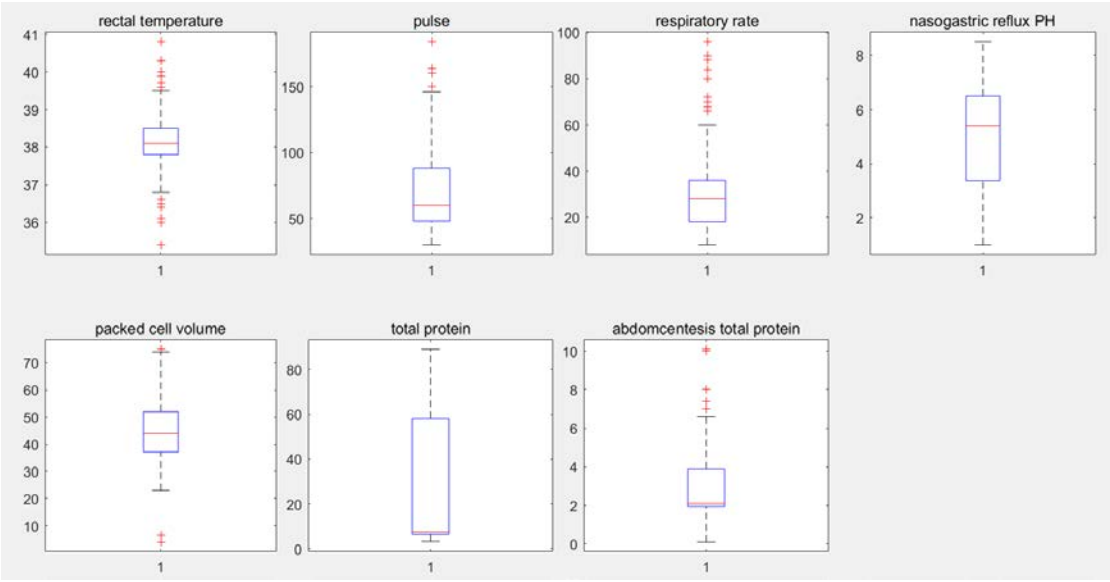


针对各数值属性，qq 图如下图所示：



通过以上 qq 图，可以看出属性“rectal temperature”与“packed cell volume” 拟合正态分布较好。

2. 绘制盒图，对离群值进行识别。



可以看出 rectal temperature 属性和 respiratory rate 属性相比较于其他属性的盒图具有较多的离群值。

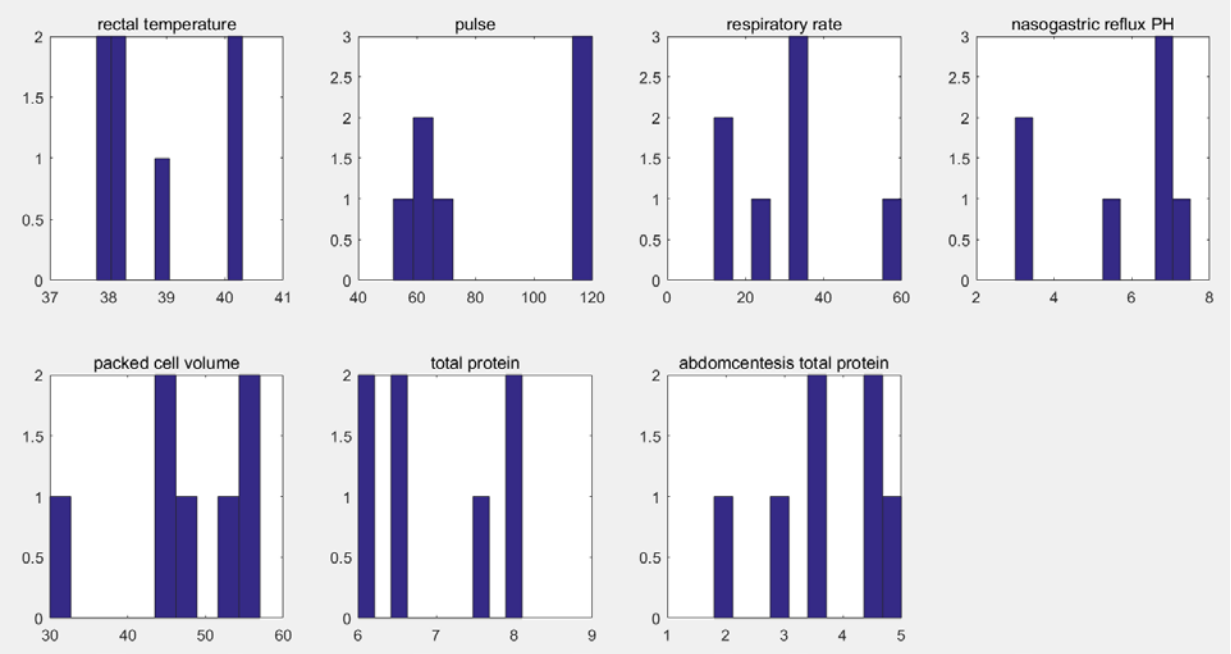
三、数据缺失的处理

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：

- 剔除缺失数据
- 按最高频率值：对于某一属性下的缺失，使用该属性的最高频率值代替缺失的数据；
- 按属性填补：计算两个属性的相关性，相关性越大表明可以根据另一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；
- 按相似性填补：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值的样本。

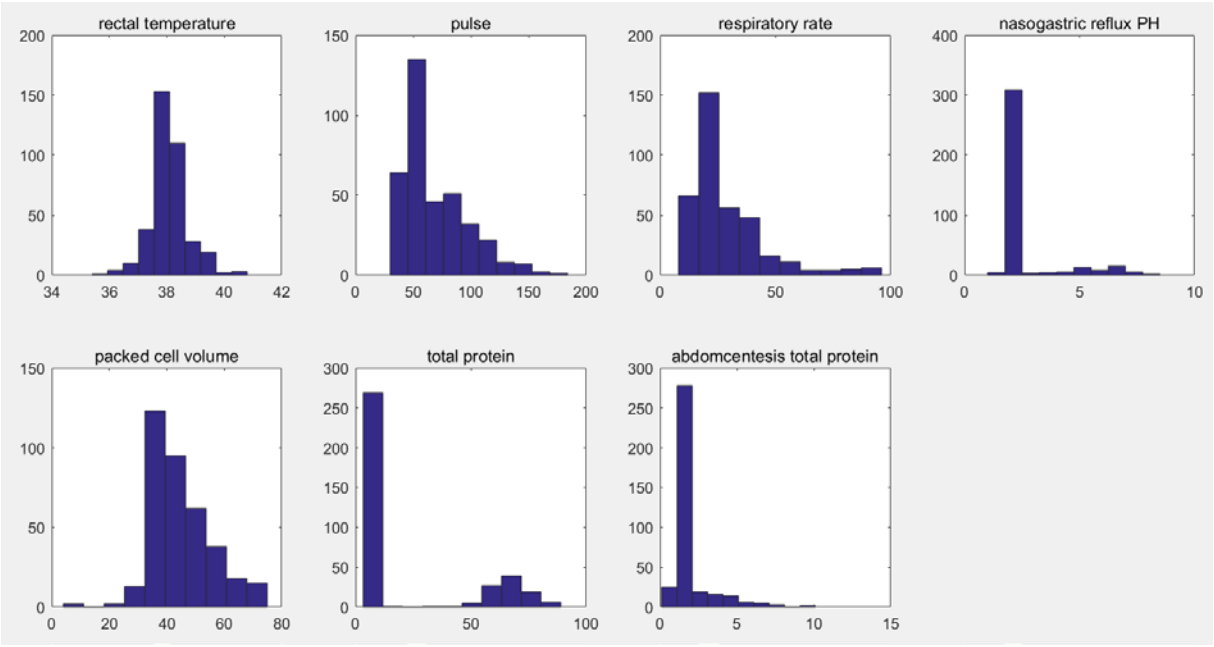
1. 将缺失部分剔除



经过删除，该数据集仅剩 7 条数据，用词方法对缺失值进行处理效果并不好。

2. 用最高频率值来填补缺失值

根据直方图，用最高频率值填补缺失值后，变化不明显。



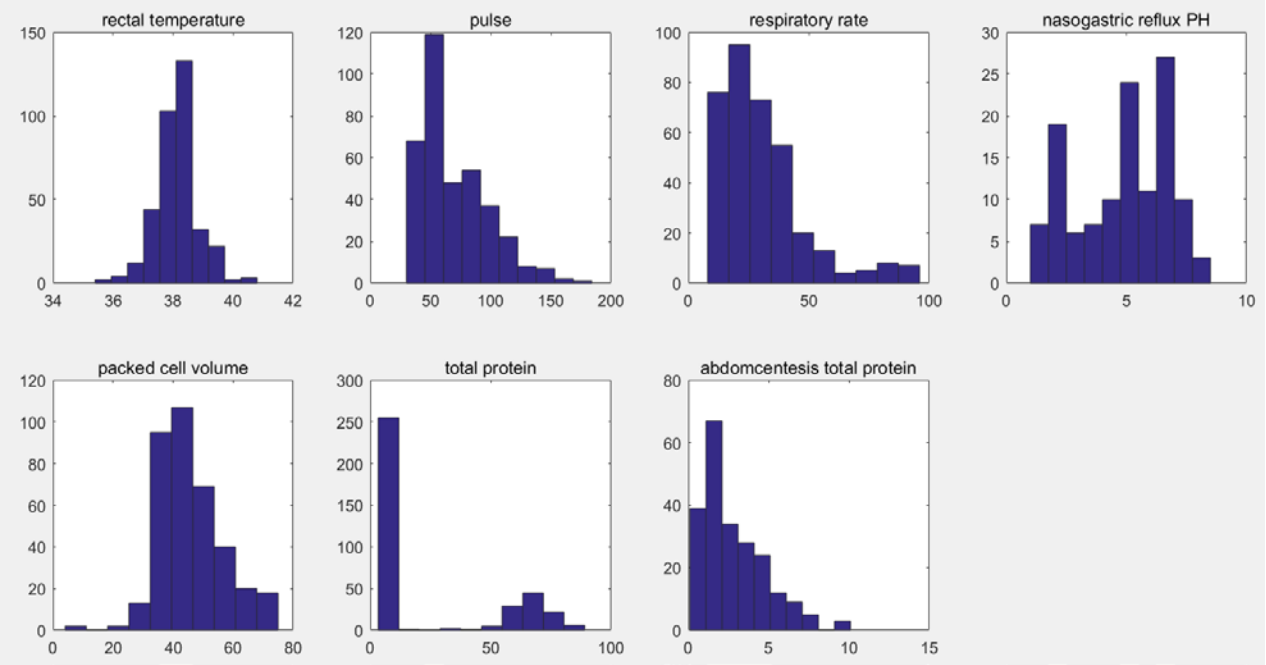
3. 通过属性的相关关系来填补缺失值

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	1.0000	0.6759	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	-0.0499	0.0012	0.0063	0.1819	-0.0990
3	NaN	0.6759	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	-0.1136	0.1304	-0.0562	-0.0194	-0.1367
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
10	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
12	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
14	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
15	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
16	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
17	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
18	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
19	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
20	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
21	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
22	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
23	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
24	NaN	-0.0499	-0.1136	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0000	-0.2515	-0.0843	-0.0400	0.0575
25	NaN	0.0012	0.1304	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	-0.2515	1.0000	-0.0024	0.0054	-0.0241
26	NaN	0.0063	-0.0562	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	-0.0843	-0.0024	1	0.2460	0.1081
27	NaN	0.1819	-0.0194	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	-0.0400	0.0054	0.2460	1	0.0372
28	NaN	-0.0990	-0.1367	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0575	-0.0241	0.1081	0.0372	1.0000

有上述结果可以看出，数值属性之间的相关性均不高。

4. 通过数据对象之间的相似性来填补缺失值

数据相似性结果见“attribute_similarity.txt”，根据数据对象之间的相似性进行填补，结果如下所示：


rectal temperature
pulse
respiratory rate
nasogastric reflux PH
packed cell volume
total protein
abdomcentesis total protein