数据挖掘——马的疝病分析

2120161061 王雨佳

一、数据摘要

为方便处理，首先将数据中确实数据“?”统一替换为“NaN”。

1. 对标称属性，给出每个可能取值的频数。

针对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，随机选择几个属性（surgery/Age/abdomen/cp_data）进行结果展示，结果如下：

>> tabulate(surgery)

Value	Count
1	214
2	152

>> tabulate(cp_data)

Value	Count
1	124
2	244

>> tabulate(abdomen)

Value	Count
1	31
2	24
3	19
4	55
5	96

>> tabulate(Age)

Value	Count
1	340
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	28

2. 数值属性，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：

surgery
max: 2
min: 1
mean: 1.4153
median: 1
quartile-1/4: 1
quartile-3/4: 2
missing: 2

Age
max: 9
min: 1
mean: 1.6087
median: 1
quartile-1/4: 1
quartile-3/4: 1
missing: 0

Hospital_Number
max: 5305629
min: 514279
mean: 1112333.8614
median: 530299
quartile-1/4: 528911.5
quartile-3/4: 534736
missing: 0

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

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

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

temperature_of_extremities
max: 4
min: 1
mean: 2.3564
median: 3
quartile-1/4: 1
quartile-3/4: 3
missing: 65

peripheral_pulse
max: 4
min: 1
mean: 1.9614
median: 1
quartile-1/4: 1
quartile-3/4: 3
missing: 83

mucous_membranes
max: 6
min: 1
mean: 2.8344
median: 3
quartile-1/4: 1
quartile-3/4: 4
missing: 48

capillary_refill_time
max: 3
min: 1
mean: 1.303
median: 1
quartile-1/4: 1
quartile-3/4: 2
missing: 38

abdominal_distension
max: 4
min: 1
mean: 2.2244
median: 2
quartile-1/4: 1
quartile-3/4: 3
missing: 65

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

pain
max: 5
min: 1
mean: 2.9082
median: 3
quartile-1/4: 2
quartile-3/4: 4
missing: 63

nasogastric_tube
max: 3
min: 1
mean: 1.7384
median: 2
quartile-1/4: 1
quartile-3/4: 2
missing: 131

rectal_examination
max: 4
min: 1
mean: 2.7792
median: 3
quartile-1/4: 1
quartile-3/4: 4
missing: 128

peristalsis
max: 4
min: 1
mean: 2.9082
median: 3
quartile-1/4: 3
quartile-3/4: 4
missing: 52

nasogastric_reflux
max: 3
min: 1
mean: 1.6085
median: 1
quartile-1/4: 1
quartile-3/4: 2
missing: 133

abdomen
max: 5
min: 1
mean: 3.7156
median: 4
quartile-1/4: 3
quartile-3/4: 5
missing: 143

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

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

lesion1
max: 41110
min: 0
mean: 3650.8342
median: 3025
quartile-1/4: 2111.5
quartile-3/4: 3209
missing: 0

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

outcome
max: 3
min: 1
mean: 1.5273
median: 1
quartile-1/4: 1
quartile-3/4: 2
missing: 2

lesion2
max: 7111
min: 0
mean: 96.9728
median: 0
quartile-1/4: 0
quartile-3/4: 0
missing: 0

abdominocentesis_appearance
max: 3
min: 1
mean: 2.046
median: 2
quartile-1/4: 1
quartile-3/4: 3
missing: 194

surgical_lesion
max: 2
min: 1
mean: 1.3696
median: 1
quartile-1/4: 1
quartile-3/4: 2
missing: 0

lesion3
max: 2209
min: 0
mean: 6.0027
median: 0
quartile-1/4: 0
quartile-3/4: 0
missing: 0

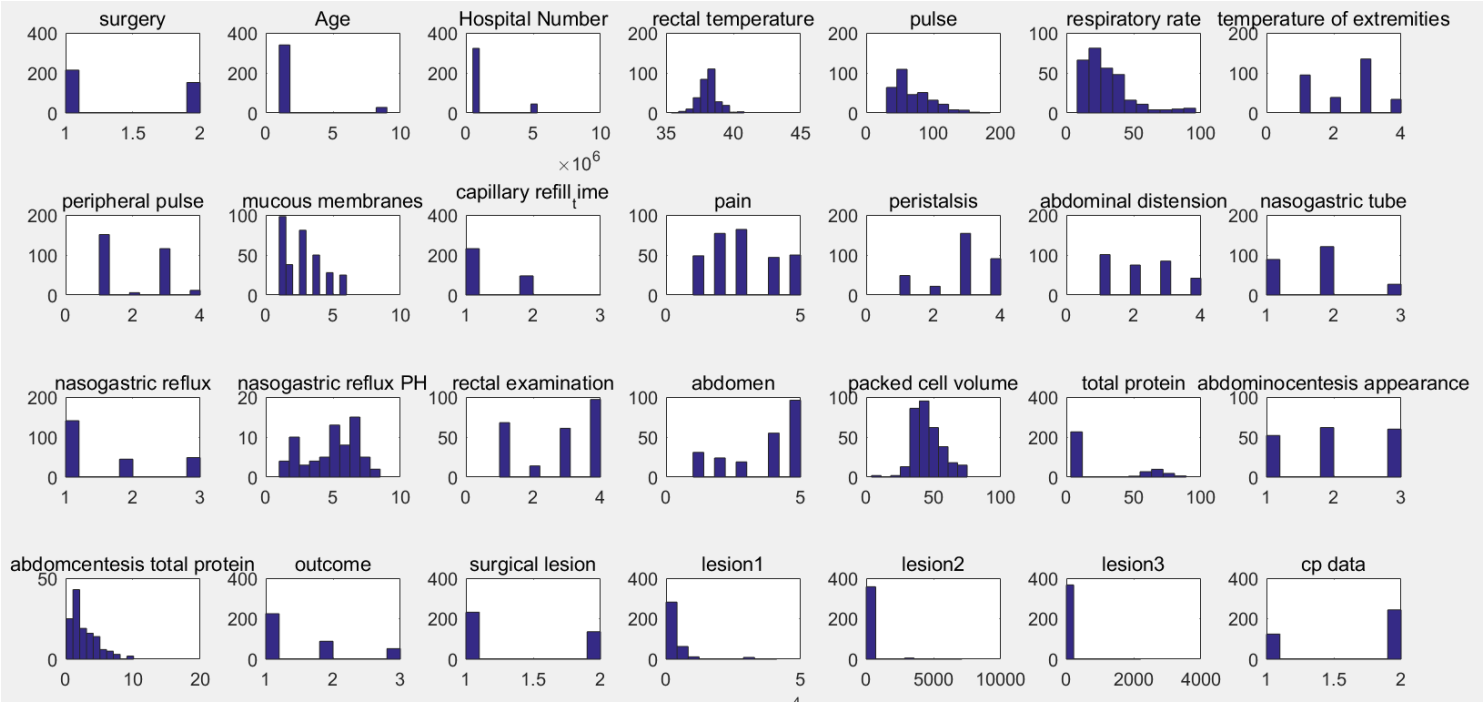
cp_data
max: 2
min: 1
mean: 1.663
median: 2
quartile-1/4: 1
quartile-3/4: 2
missing: 0

二、数据可视化

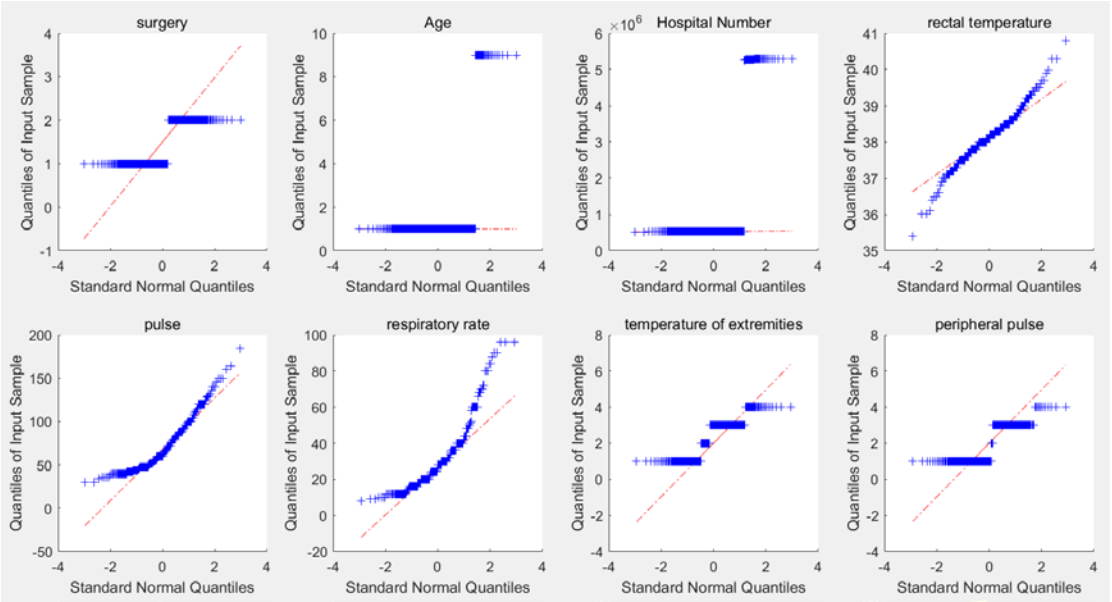
针对数值属性，

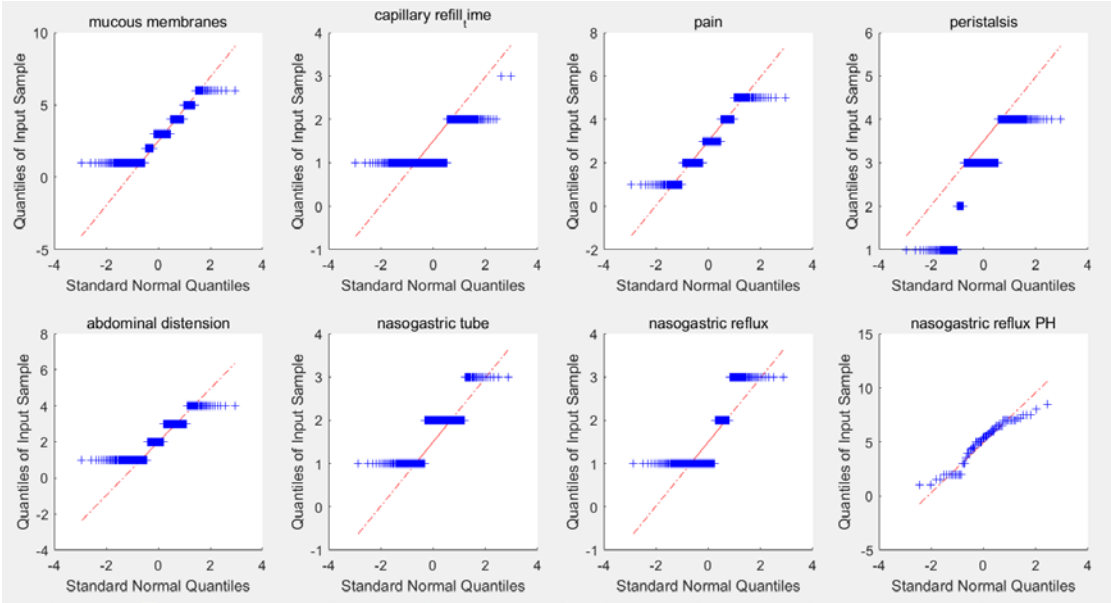
1. 绘制直方图，如 mxPH，用 qq 图检验其分布是否为正态分布。

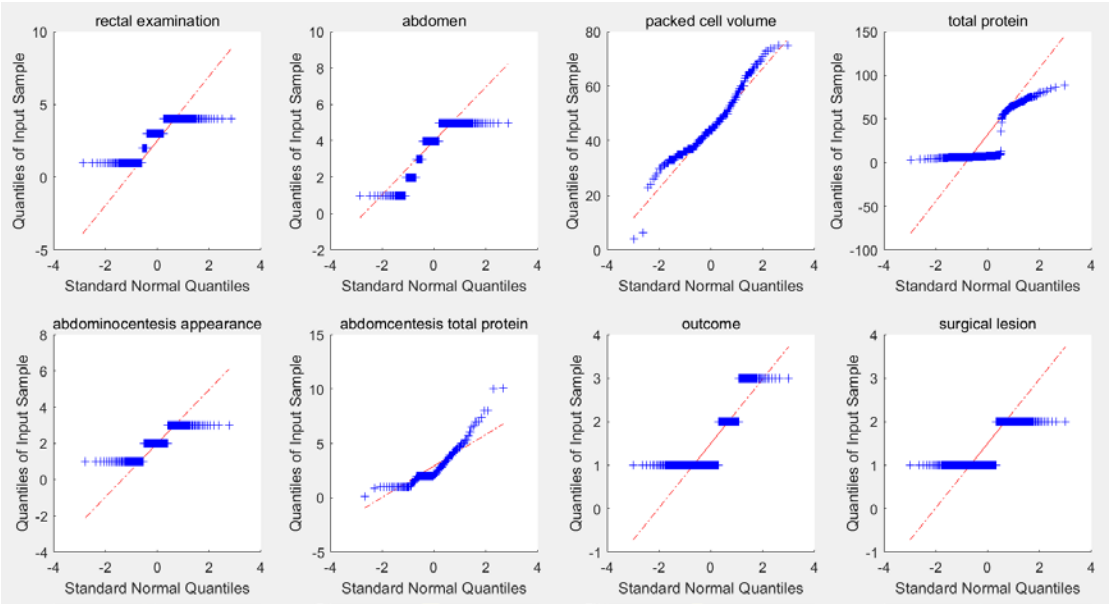
针对各标称属性，直方图如下所示：

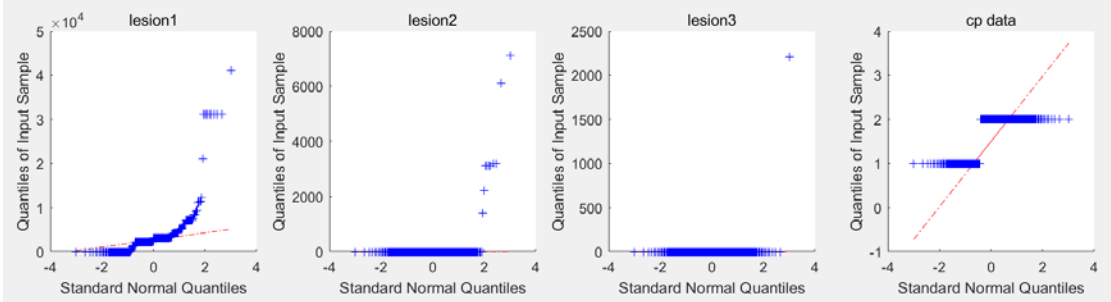


针对各标称属性，qq 图如下图所示：



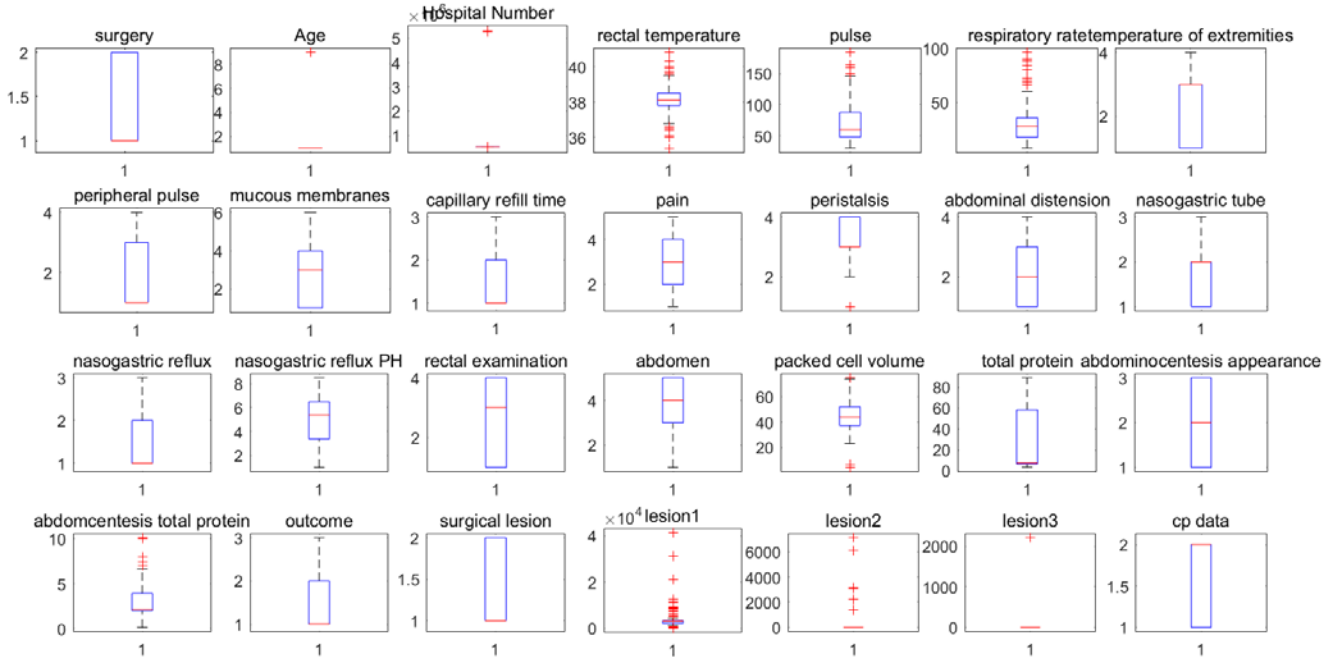






通过以上 qq 图，可以看出属性“packed cell volume”服从正态分布。

2. 绘制盒图，对离群值进行识别。

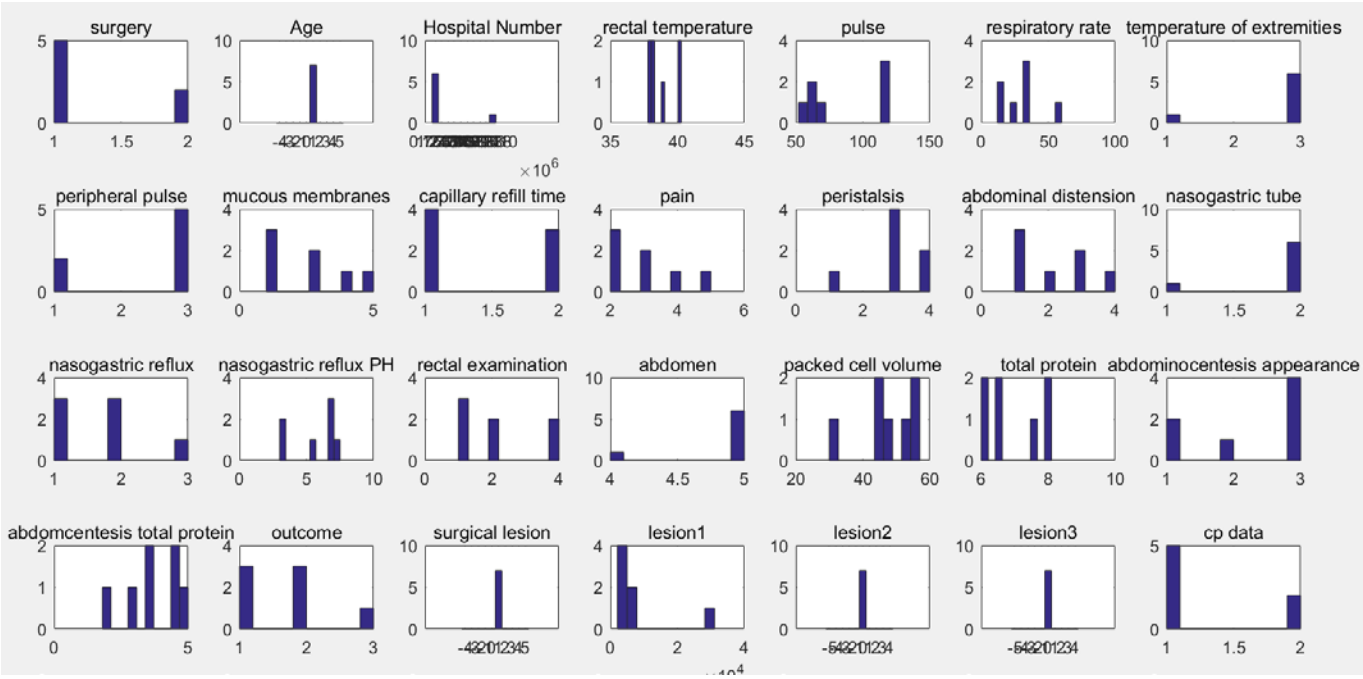


三、数据缺失的处理

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。

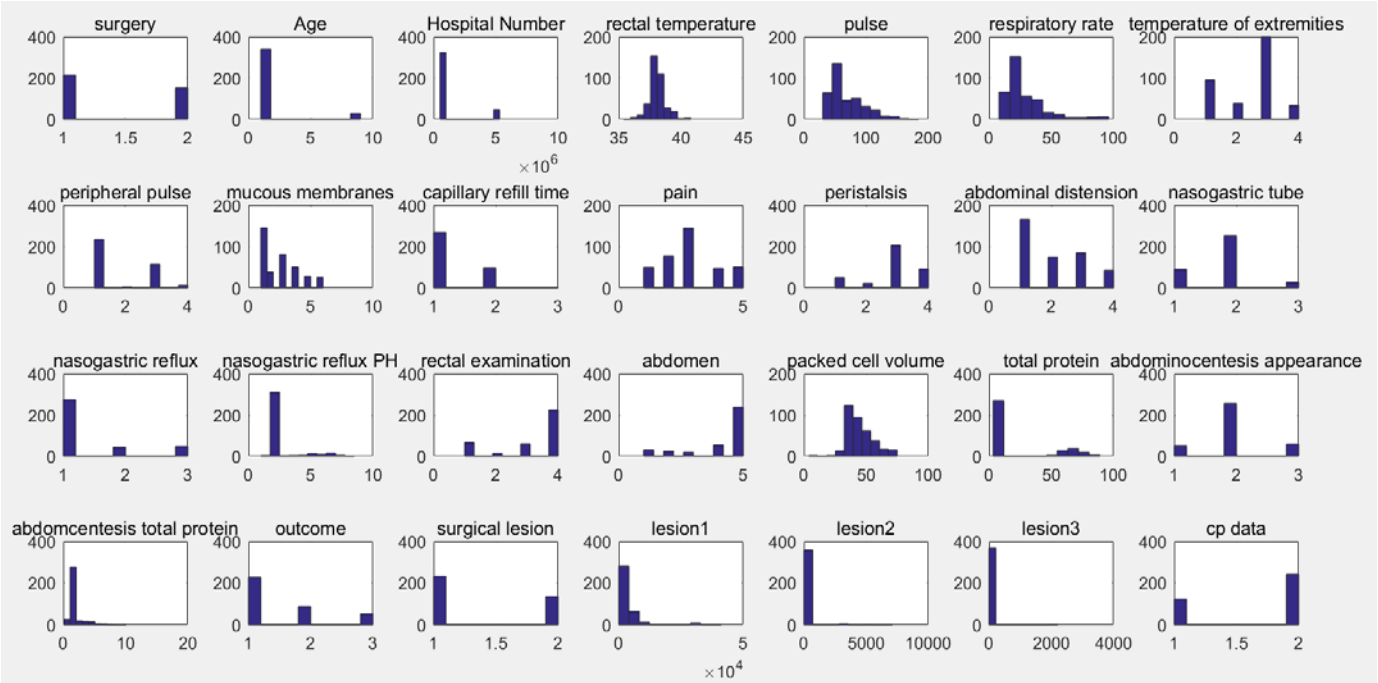
分别使用下列四种策略对缺失值进行处理：

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。

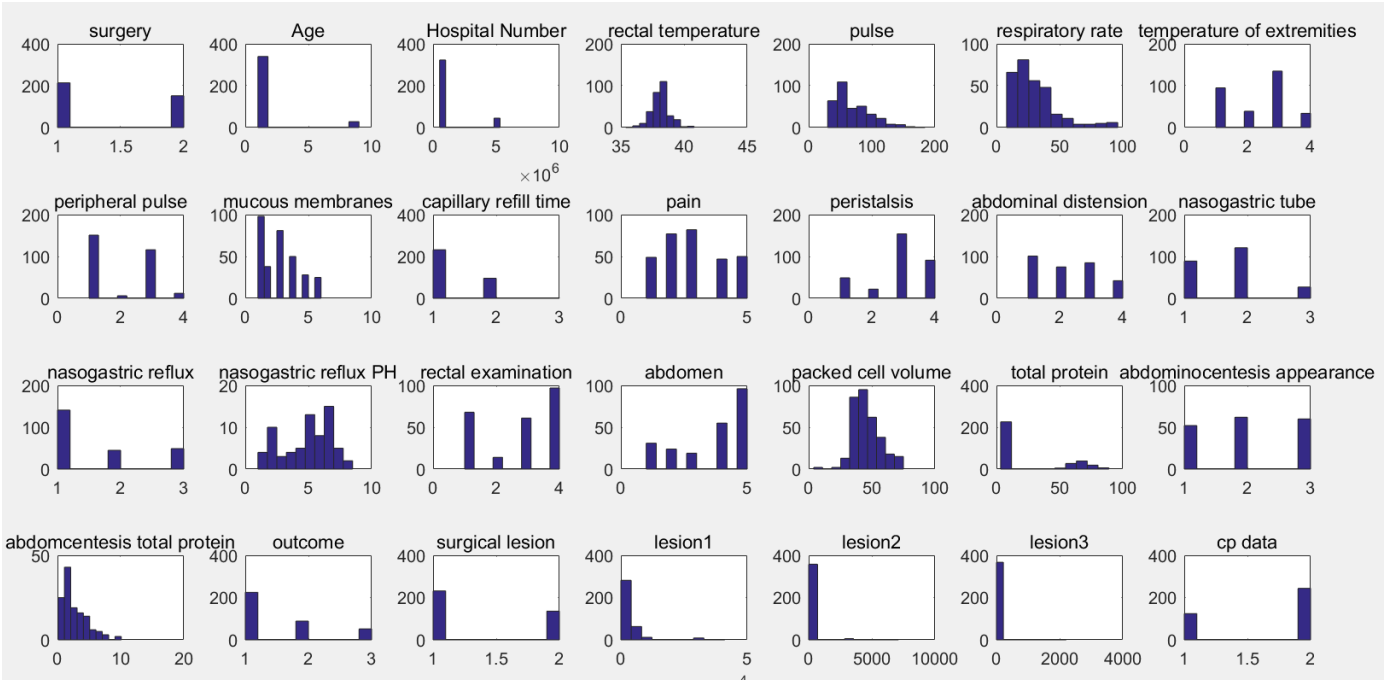
3. 通过属性的相关关系来填补缺失值

	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，属性“Age”与“Hospital Number”显著相关（0.6759），“surgical lesion”与“lesion1”微相关（-0.2515），“lesion2”与“lesion3”微相关（0.2460）。

Hospital Number = +(4.9647e+05) * Age + (3.1367e+05)

Age = +(9.2019e-07) * Hospital Number + (0.58514)



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：