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

```
surgery 的频数为：
 1.0    214
2.0    152
dtype: int64

 Age  的频数为：
 1    340
9     28
dtype: int64
```

```
Hospital Number 的频数为：
 530670     2
530526      2
5279822     2
529461      2
528151      2
5274919     2
528729      2
527544      2
527916      2
529424      2
529796      2
528931      2
533815      2
530239      2
528996      2
528904      2
530693      2
528469      2
5291329     2
528890      2
528926      2
532349      2
535208      1
528268      1
530431      1
521399      1
528047      1
529615      1
530612      1
530101      1
           ..
529567      1
530294      1
535415      1
529272      1
530297      1
5294369     1
530301      1
535031      1
529766      1
530276      1
535364      1
535392      1
530033      1
5275212     1
529736      1
534857      1
530251      1
533836      1
530254      1
530255      1
528999      1
527698      1
533750      1
535381      1
528214      1
533847      1
527706      1
527709      1
535338      1
530576      1
dtype: int64
```

```
rectal temperature 的频数为：
 38.0    34
38.3    23
38.2    23
37.8    21
38.5    21
37.5    16
38.1    15
38.4    14
38.6    14
37.6    10
37.7    10
37.9     9
38.7     9
37.2     7
37.3     6
39.0     6
38.8     6
39.2     5
37.1     5
37.0     4
39.3     4
39.5     4
38.9     4
37.4     4
39.1     3
39.4     3
36.5     2
36.0     2
39.7     2
36.6     2
40.3     2
39.9     1
35.4     1
40.0     1
36.9     1
39.6     1
36.4     1
36.8     1
40.8     1
36.1     1
dtype: int64
```

```
pulse  的频数为：
 48.0     35
60.0      33
40.0      21
44.0      16
88.0      15
42.0      15
52.0      14
72.0      13
100.0     13
120.0     12
84.0      11
80.0      11
54.0       9
56.0       9
96.0       8
64.0       8
66.0       7
50.0       6
104.0      5
68.0       5
92.0       5
76.0       5
78.0       5
70.0       4
45.0       4
90.0       4
36.0       4
150.0      3
86.0       3
112.0      3
108.0      3
114.0      3
140.0      3
30.0       2
82.0       2
65.0       2
130.0      2
75.0       2
124.0      2
164.0      1
98.0       1
184.0      1
38.0       1
49.0       1
146.0      1
160.0      1
132.0      1
46.0       1
129.0      1
128.0      1
55.0       1
110.0      1
136.0      1
34.0       1
dtype: int64
```

```
respiratory rate  的频数为：
 20.0    35
24.0    30
12.0    27
16.0    26
30.0    26
40.0    22
36.0    20
28.0    16
32.0    12
18.0    10
60.0    10
48.0     6
44.0     5
14.0     4
10.0     4
96.0     3
35.0     3
68.0     3
80.0     3
42.0     3
84.0     2
70.0     2
21.0     2
90.0     2
72.0     2
22.0     2
51.0     2
9.0      2
50.0     2
23.0     1
15.0     1
34.0     1
25.0     1
8.0      1
52.0     1
58.0     1
13.0     1
66.0     1
26.0     1
88.0     1
dtype: int64
```

```
 temperature of extremities 的频数为：
 3.0    135
1.0     95
2.0     39
4.0     34
dtype: int64

peripheral pulse 的频数为：
 1.0    151
3.0    116
4.0     12
2.0      6
dtype: int64

mucous membranes 的频数为：
 1.0    98
3.0    81
4.0    50
2.0    38
5.0    28
6.0    25
dtype: int64

capillary refill time  的频数为：
 1.0    232
2.0     96
3.0      2
dtype: int64

 pain 的频数为：
 3.0    82
2.0    77
5.0    50
1.0    49
4.0    47
dtype: int64

peristalsis  的频数为：
 3.0    154
4.0     91
1.0     49
2.0     22
dtype: int64

 abdominal distension 的频数为：
 1.0    101
3.0     85
2.0     75
4.0     42
dtype: int64
```

```
nasogastric tube  的频数为：
 2.0    121
1.0     89
3.0     27
dtype: int64

nasogastric reflux  的频数为：
 1.0    141
3.0     49
2.0     45
dtype: int64

 nasogastric reflux PH  的频数为：
 2.0    10
7.0     9
5.0     7
6.5     6
5.5     4
6.0     4
3.0     3
4.5     3
4.0     3
7.5     3
1.5     2
1.0     2
7.2     2
8.5     1
3.5     1
5.3     1
8.0     1
4.3     1
5.8     1
6.2     1
4.4     1
5.4     1
5.9     1
5.7     1
dtype: int64

 rectal examination 的频数为：
 4.0    97
1.0    68
3.0    61
2.0    14
dtype: int64
```

```
 abdomen  的频数为：
 5.0    96
4.0    55
1.0    31
2.0    24
3.0    19
dtype: int64

 packed cell volume  的频数为：
 37.0    22
45.0    17
44.0    17
43.0    17
50.0    16
35.0    16
40.0    14
36.0    13
47.0    13
48.0    11
42.0    11
33.0    10
38.0    10
41.0    10
39.0     9
60.0     9
46.0     9
54.0     8
52.0     7
49.0     7
55.0     6
57.0     6
65.0     6
53.0     6
64.0     5
34.0     5
68.0     4
59.0     4
32.0     4
30.0     3
58.0     3
66.0     3
73.0     2
74.0     2
51.0     2
31.0     2
56.0     2
63.0     2
75.0     2
69.0     2
31.5     1
4.0      1
62.0     1
28.0     1
37.5     1
72.0     1
71.0     1
67.0     1
26.0     1
24.0     1
27.0     1
23.0     1
70.0     1
6.4      1
dtype: int64
```

total protein 的频数为：
7.5   17
6.5   16
7.0   15
6.6   12
6.0   11
5.9   10
65.0   10
6.8   9
7.2   9
6.7   9
6.2   8
8.5   7
70.0   7
6.3   7
6.4   6
5.8   6
8.0   6
6.1   6
75.0   5
7.7   5
5.7   5
64.0   5
56.0   5
7.6   5
69.0   5
7.8   4
61.0   4
67.0   4
8.6   4
5.5   4
        ..
8.7   2
8.2   2
7.1   2
81.0   2
4.9   2
82.0   1
8.8   1
11.0   1
86.0   1
46.0   1
7.9   1
9.1   1
3.3   1
4.6   1
10.2   1
4.7   1
79.0   1
36.0   1
3.5   1
5.0   1
9.0   1
13.0   1
50.0   1
8.9   1
4.0   1
51.0   1
89.0   1
59.0   1
53.0   1
85.0   1
dtype: int64

```
 abdominocentesis appearance  的频数为：
 2.0    62
3.0    60
1.0    52
dtype: int64

 abdomcentesis total protein 的频数为：
 2.0     33
1.0     21
3.9      5
5.0      4
2.8      4
2.6      4
3.6      3
4.3      3
3.4      3
7.0      3
1.4      3
1.1      2
2.5      2
2.1      2
8.0      2
3.1      2
2.2      2
4.5      2
2.3      2
5.3      2
3.0      2
4.6      2
1.6      2
6.0      2
4.1      2
1.5      1
10.0     1
6.5      1
0.9      1
4.4      1
2.9      1
1.3      1
4.8      1
5.2      1
5.7      1
6.6      1
3.3      1
7.4      1
0.1      1
4.7      1
3.7      1
10.1     1
3.2      1
1.8      1
dtype: int64
```

```
outcome  的频数为：
 1.0    225
2.0     89
3.0     52
dtype: int64

surgical lesion 的频数为：
 1    232
2    136
dtype: int64

 lesion 1 的频数为：
 0        67
3111     41
3205     35
2208     23
2205     17
2209     15
4205     11
7111     10
1400     10
31110     9
2124      9
2113      8
400       7
2112      6
3209      6
2206      5
4124      5
2111      4
5400      4
3124      4
6112      4
4300      4
7209      3
3112      3
4206      3
6111      3
2207      3
5111      3
5206      2
1124      2
```

```
          ..
11124     2
3025      2
9400      2
3113      2
2300      2
8400      2
300       1
3115      1
3400      1
41110     1
5000      1
3133      1
12208     1
4122      1
4111      1
5110      1
11300     1
6209      1
9000      1
2305      1
5205      1
7400      1
1111      1
3300      1
8300      1
8405      1
4207      1
21110     1
3207      1
11400     1
dtype: int64

 lesion 2 的频数为：
 0       358
3111      3
3205      2
6112      1
7111      1
1400      1
2208      1
3112      1
dtype: int64

 lesion 3 的频数为：
 0       367
2209      1
dtype: int64

cp_data  的频数为：
 2    244
1    124
dtype: int64
```

- 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。

	quartile	missing
surgery	1.00	2.0
Age	1.00	0.0
Hospital Number	528915.25	0.0
rectal temperature	37.80	69.0
pulse	48.00	26.0
respiratory rate	18.00	71.0
temperature of extremities	1.00	65.0
peripheral pulse	1.00	83.0
mucous membranes	1.00	48.0
capillary refill time	1.00	38.0
pain	2.00	63.0
peristalsis	3.00	52.0
abdominal distension	1.00	65.0
nasogastric tube	1.00	131.0
nasogastric reflux	1.00	133.0
nasogastric reflux PH	3.50	299.0
rectal examination	1.00	128.0
abdomen	3.00	143.0
packed cell volume	37.25	37.0
total protein	6.50	43.0
abdominocentesis appearance	1.00	194.0
abdomcentesis total protein	2.00	235.0
outcome	1.00	2.0
surgical lesion	1.00	0.0
lesion 1	2111.75	0.0
lesion 2	0.00	0.0
lesion 3	0.00	0.0
cp_data	1.00	0.0

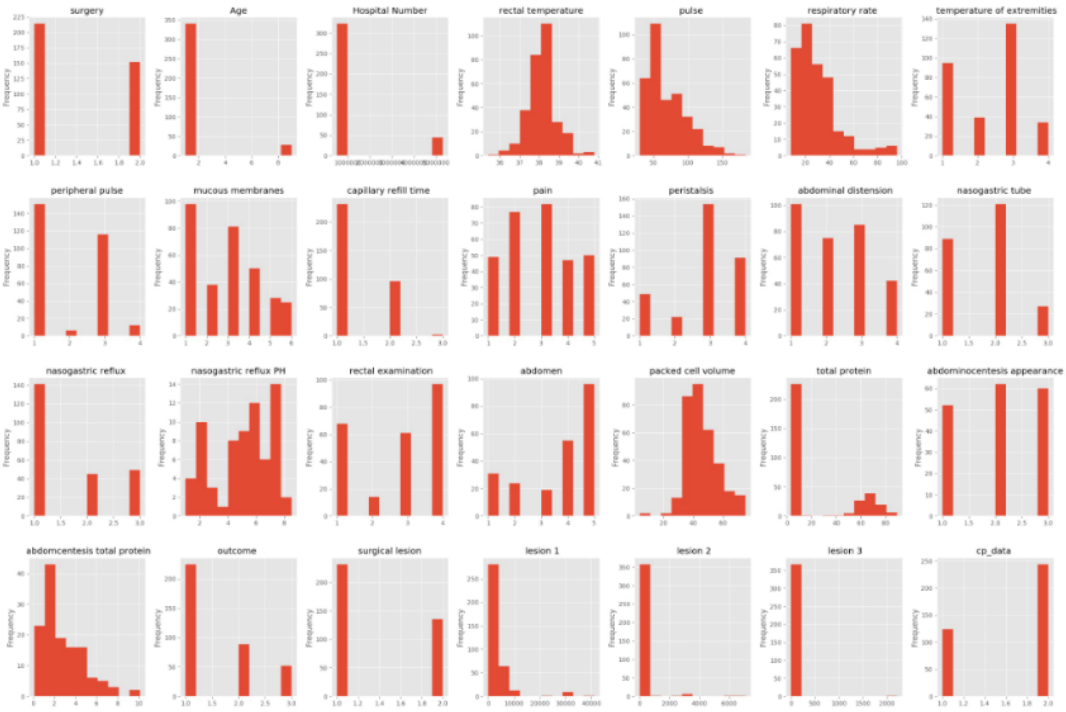
	max	min	mean	median	\
surgery	2.0	1.0	1.415301e+00	1.0	
Age	9.0	1.0	1.608696e+00	1.0	
Hospital Number	5305629.0	514279.0	1.112334e+06	530299.0	
rectal temperature	40.8	35.4	3.813445e+01	38.1	
pulse	184.0	30.0	7.075731e+01	60.0	
respiratory rate	96.0	8.0	3.052189e+01	28.0	
temperature of extremities	4.0	1.0	2.356436e+00	3.0	
peripheral pulse	4.0	1.0	1.961404e+00	1.0	
mucous membranes	6.0	1.0	2.834375e+00	3.0	
capillary refill time	3.0	1.0	1.303030e+00	1.0	
pain	5.0	1.0	2.908197e+00	3.0	
peristalsis	4.0	1.0	2.908228e+00	3.0	
abdominal distension	4.0	1.0	2.224422e+00	2.0	
nasogastric tube	3.0	1.0	1.738397e+00	2.0	
nasogastric reflux	3.0	1.0	1.608511e+00	1.0	
nasogastric reflux PH	8.5	1.0	4.962319e+00	5.4	
rectal examination	4.0	1.0	2.779167e+00	3.0	
abdomen	5.0	1.0	3.715556e+00	4.0	
packed cell volume	75.0	4.0	4.565680e+01	44.0	
total protein	89.0	3.3	2.477108e+01	7.5	
abdominocentesis appearance	3.0	1.0	2.045977e+00	2.0	
abdomcentesis total protein	10.1	0.1	2.948120e+00	2.1	
outcome	3.0	1.0	1.527322e+00	1.0	
surgical lesion	2.0	1.0	1.369565e+00	1.0	
lesion 1	41110.0	0.0	3.650834e+03	3025.0	
lesion 2	7111.0	0.0	9.697283e+01	0.0	
lesion 3	2209.0	0.0	6.002717e+00	0.0	
cp_data	2.0	1.0	1.663043e+00	2.0	

3.1.2 数据的可视化

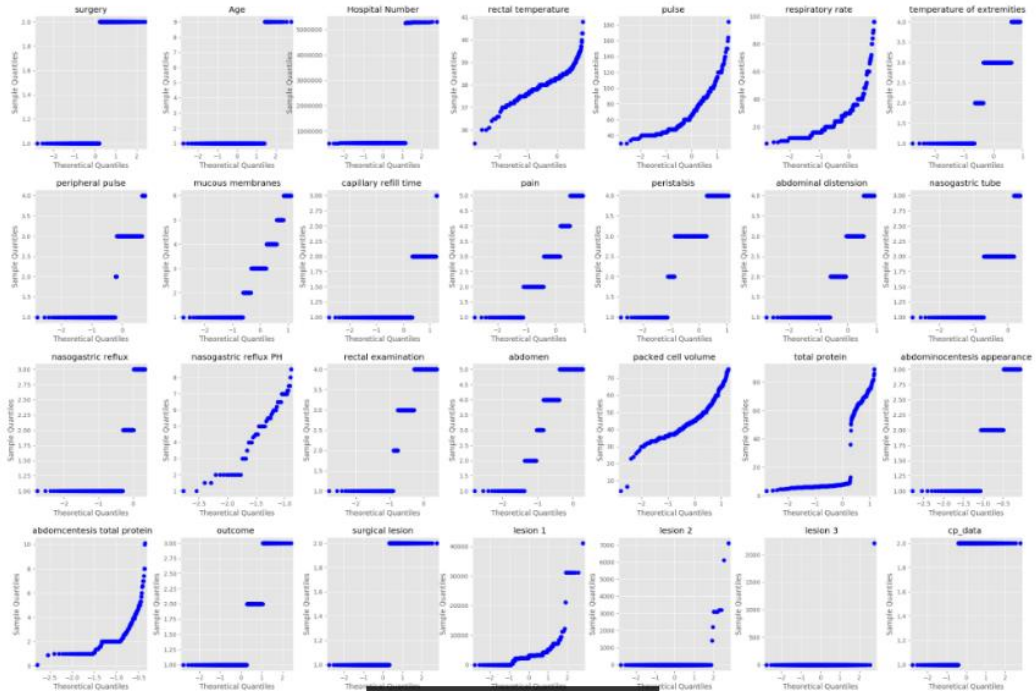
针对数值属性：

- 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：

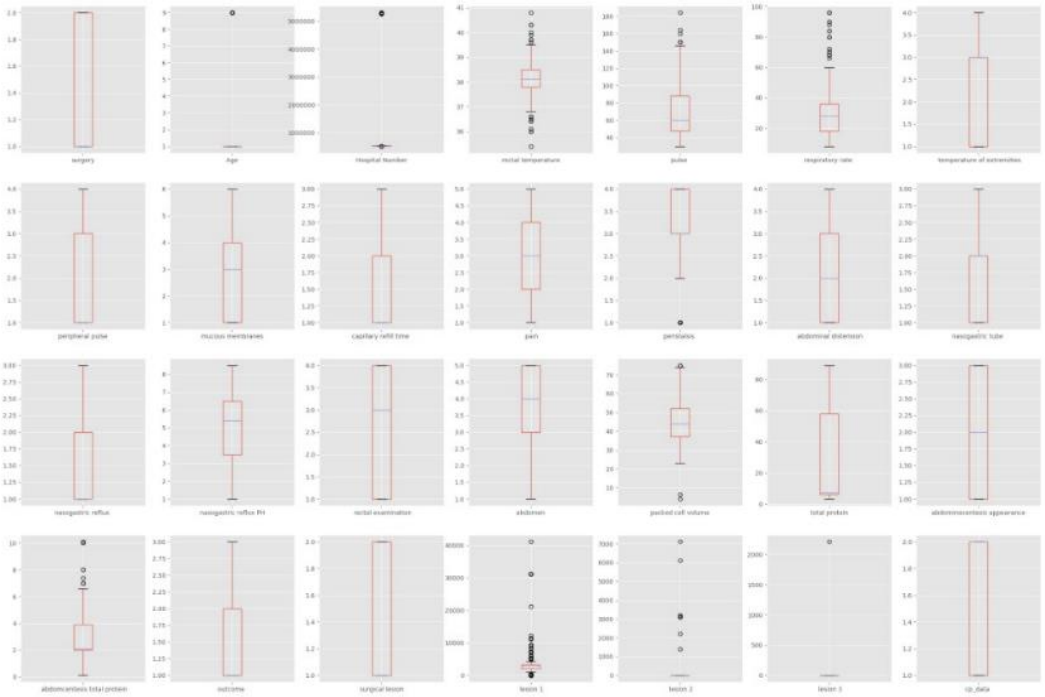


qq 图如下所示：



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



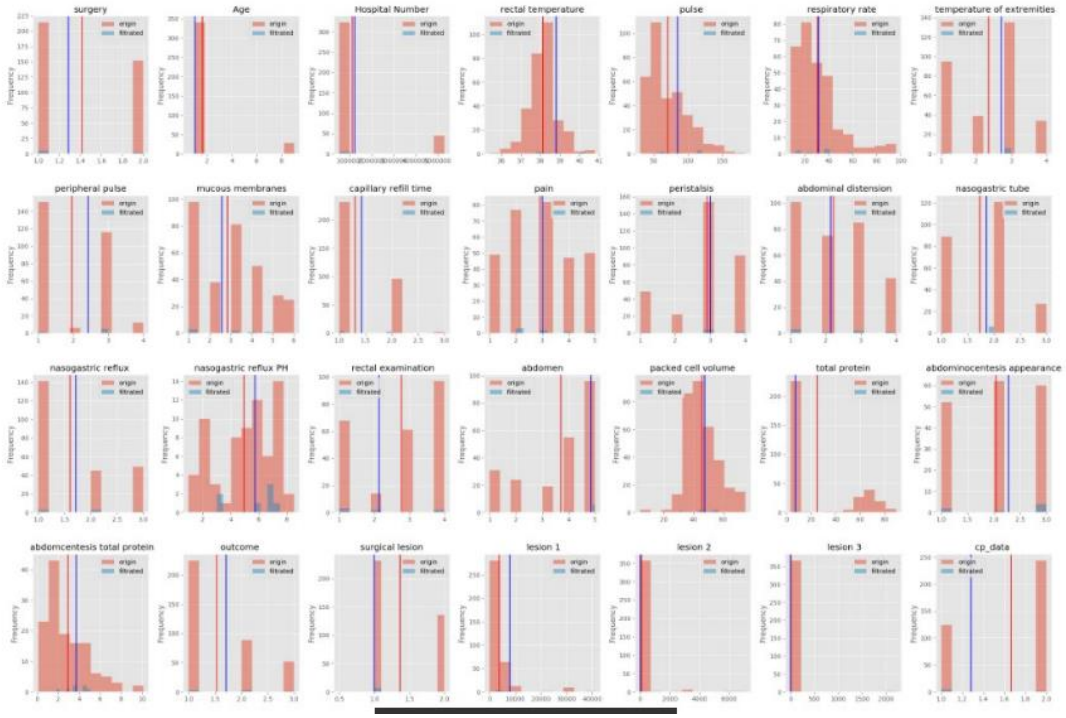
根据各个属性的盒图可以看出，属性 Age，Hospital Number，rectal temperature，pulse，respiratory rate，peristalsis，packed cell volume，abdomcentesis total protein，lesion1，lesion2，lesion3 存在离群值。

3.2 数据缺失的处理

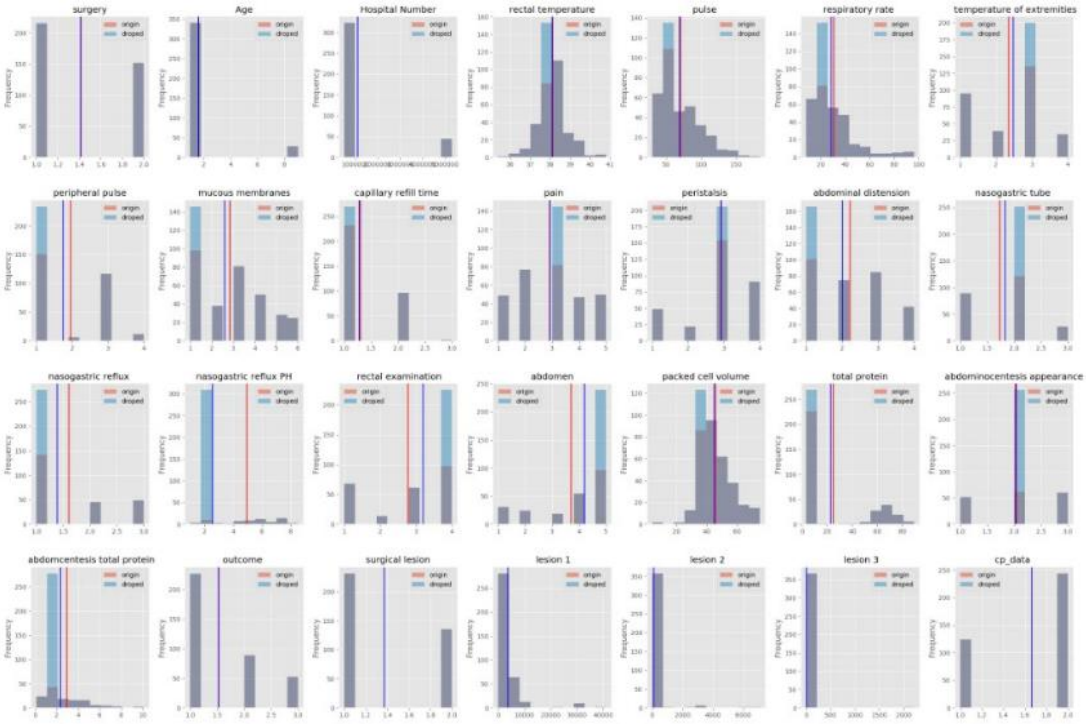
数据集中有 30%的值是缺失的，因此需要先处理数据中的缺失值。

分别使用下列四种策略对缺失值进行处理:

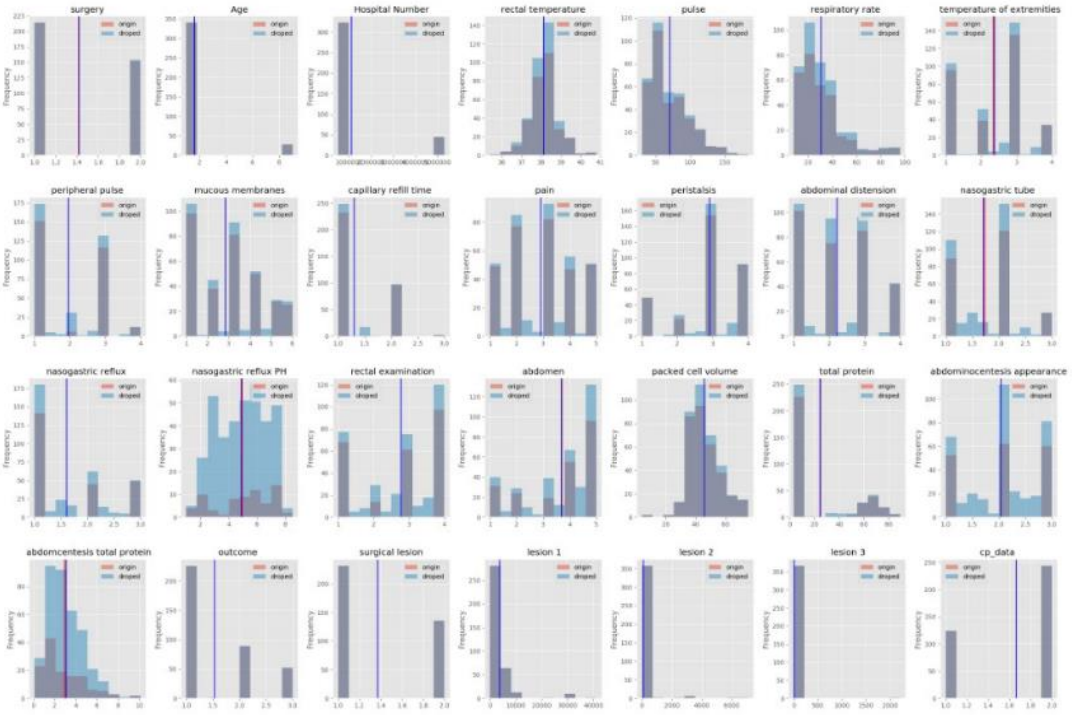
- 将缺失部分剔除



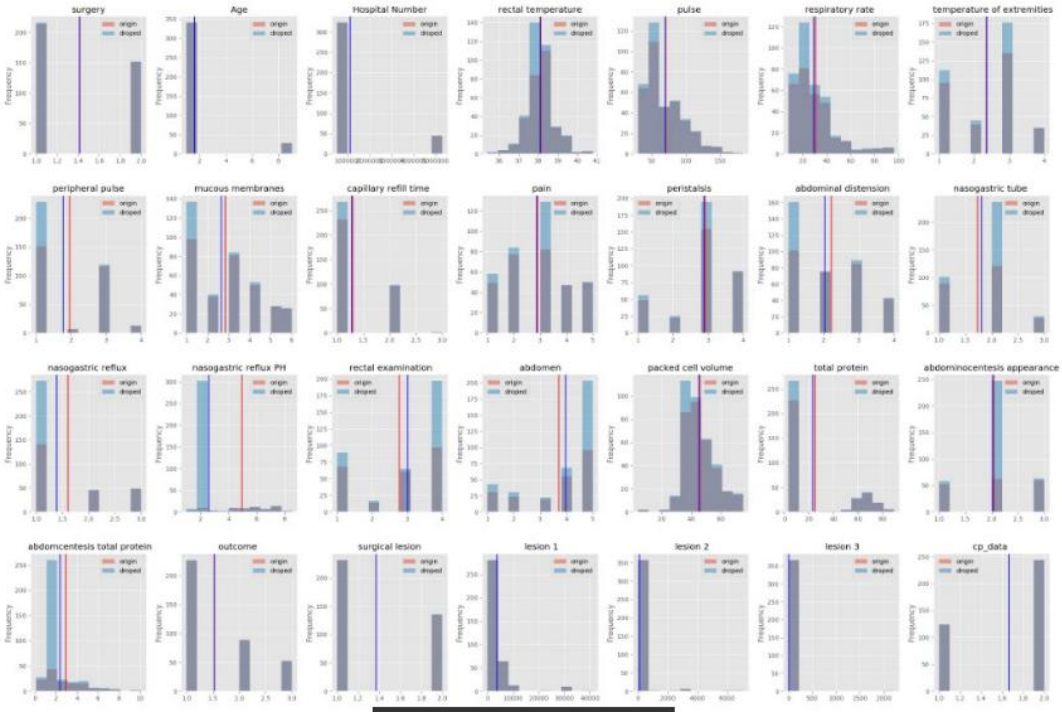
- 用最高频率值来填补缺失值



- 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


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
origin
droped
Frequency