

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：余非梧

学号：2120161074

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 300 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      180  
2.0      119  
dtype: int64  
  
Age 的频数为:  
1        276  
9         24  
dtype: int64
```

Hospital Number 的频数为:

529796	2
529424	2
5279822	2
527916	2
530526	2
528996	2
528729	2
528151	2
529461	2
528931	2
528890	2
528469	2
530693	2
532349	2
528904	2
527544	2
533697	1
533696	1
533736	1
534719	1
530101	1
529399	1
533692	1
528570	1
527957	1
534157	1
521399	1
529045	1
530612	1
528047	1
..	
533887	1
535031	1
530301	1
5294369	1
530297	1
529272	1
535415	1
530294	1
534899	1
529777	1
535407	1
522979	1
528743	1
534885	1
534857	1
530276	1
535392	1
535338	1
527709	1
527706	1
533847	1
528214	1
535381	1
533750	1
527698	1
530255	1
530254	1
533836	1
530251	1
535043	1

dtype: int64

respiratory rate 的频数为:

20.0	28
24.0	27
16.0	22
12.0	19
30.0	19
40.0	17
36.0	16
28.0	13
32.0	11
18.0	8
48.0	6
14.0	4
60.0	4
35.0	3
68.0	3
10.0	3
80.0	3
44.0	3
42.0	3
90.0	2
84.0	2
70.0	2
21.0	2
96.0	2
22.0	2
51.0	2
9.0	2
50.0	2
72.0	1
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	109
1.0	78
2.0	30
4.0	27

dtype: int64

peripheral pulse 的频数为:

1.0	115
3.0	103
4.0	8
2.0	5

dtype: int64

mucous membranes 的频数为:

1.0	79
3.0	58
4.0	41
2.0	30
5.0	25
6.0	20

dtype: int64

capillary refill time 的频数为:

1.0	188
2.0	78
3.0	2

dtype: int64

pain 的频数为:

3.0	67
2.0	59
5.0	42
4.0	39
1.0	38

dtype: int64

peristalsis 的频数为:

3.0	128
4.0	73
1.0	39
2.0	16

dtype: int64

abdominal distension 的频数为:

1.0	76
3.0	65
2.0	65
4.0	38

dtype: int64

nasogastric tube 的频数为:

2.0	102
1.0	71
3.0	23

dtype: int64

nasogastric reflux 的频数为:

1.0	120
3.0	39
2.0	35

dtype: int64

nasogastric reflux PH 的频数为:

2.0	9
7.0	8
6.5	5
5.0	4
5.5	4
3.0	3
4.5	3
4.0	3
1.0	2
7.5	2
6.0	1
5.7	1
5.3	1
3.5	1
4.3	1
7.2	1
6.2	1
4.4	1
5.4	1
1.5	1

dtype: int64

rectal examination 的频数为:

4.0	79
1.0	57
3.0	49
2.0	13

dtype: int64

abdomen 的频数为:

5.0	79
4.0	43
1.0	28
2.0	19
3.0	13

dtype: int64

```
abdominocentesis appearance 的频数为:  
2.0    48  
3.0    46  
1.0    41  
dtype: int64
```

```
abdomcentesis total protein 的频数为:  
2.0    26  
1.0    17  
2.6     4  
3.9     4  
5.0     3  
3.6     3  
2.8     3  
7.0     3  
3.4     3  
3.0     2  
4.1     2  
2.3     2  
1.6     2  
6.0     2  
4.3     2  
4.5     2  
5.3     2  
2.5     1  
10.0    1  
8.0     1  
1.5     1  
2.9     1  
4.4     1  
1.3     1  
4.8     1  
5.2     1  
2.1     1  
2.2     1  
1.4     1  
3.3     1  
7.4     1  
0.1     1  
4.7     1  
3.7     1  
10.1    1  
3.2     1  
6.6     1  
dtype: int64
```

```
outcome 的频数为:  
1.0    178  
2.0     77  
3.0     44  
dtype: int64
```

```
surgical lesion 的频数为:  
1     191  
2     109  
dtype: int64
```

```

lesion 1 的频数
0          56
3111       33
3205       29
2208       20
2205       13
4205       11
2209       11
2124        9
1400        8
31110       7
7111        7
2113        6
2112        5
400         5
3209        4
4300        4
2206        4
5400        4
3112        3
4124        3
2111        3
2207        3
7209        3
5206        2
5124        2
3124        2
5111        2
9400        2
6111        2
2322        2
..
3025        2
8400        2
6112        2
11300       1
4122        1
7113        1
6209        1
3115        1
5000        1
3133        1
4111        1
3400        1
300         1
12208       1
9000        1
5205        1
1111        1
1124        1
8300        1
2305        1
4206        1
4207        1
21110       1
2300        1
3207        1
11400       1
7400        1
3113        1
3300        1
41110       1
dtype: int64

```

```

lesion 2 的频数为:
0      293
3111    3
6112    1
7111    1
1400    1
3112    1
dtype: int64

lesion 3 的频数为:
0      299
2209    1
dtype: int64

cp_data 的频数为:
2      201
1       99
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。

	max	min	mean	median \
surgery	2.0	1.0	1.397993e+00	1.00
Age	9.0	1.0	1.640000e+00	1.00
Hospital Number	5305629.0	518476.0	1.085889e+06	530305.50
rectal temperature	40.8	35.4	3.816792e+01	38.20
pulse	184.0	30.0	7.191304e+01	64.00
respiratory rate	96.0	8.0	3.041736e+01	24.50
temperature of extremities	4.0	1.0	2.348361e+00	3.00
peripheral pulse	4.0	1.0	2.017316e+00	2.00
mucous membranes	6.0	1.0	2.853755e+00	3.00
capillary refill time	3.0	1.0	1.305970e+00	1.00
pain	5.0	1.0	2.951020e+00	3.00
peristalsis	4.0	1.0	2.917969e+00	3.00
abdominal distension	4.0	1.0	2.266393e+00	2.00
nasogastric tube	3.0	1.0	1.755102e+00	2.00
nasogastric reflux	3.0	1.0	1.582474e+00	1.00
nasogastric reflux PH	7.5	1.0	4.707547e+00	5.00
rectal examination	4.0	1.0	2.757576e+00	3.00
abdomen	5.0	1.0	3.692308e+00	4.00
packed cell volume	75.0	23.0	4.629520e+01	45.00
total protein	89.0	3.3	2.445693e+01	7.50
abdominocentesis appearance	3.0	1.0	2.037037e+00	2.00
abdomcentesis total protein	10.1	0.1	3.019608e+00	2.25
outcome	3.0	1.0	1.551839e+00	1.00
surgical lesion	2.0	1.0	1.363333e+00	1.00
lesion 1	41110.0	0.0	3.657880e+03	2673.50
lesion 2	7111.0	0.0	9.022667e+01	0.00
lesion 3	2209.0	0.0	7.363333e+00	0.00
cp_data	2.0	1.0	1.670000e+00	2.00

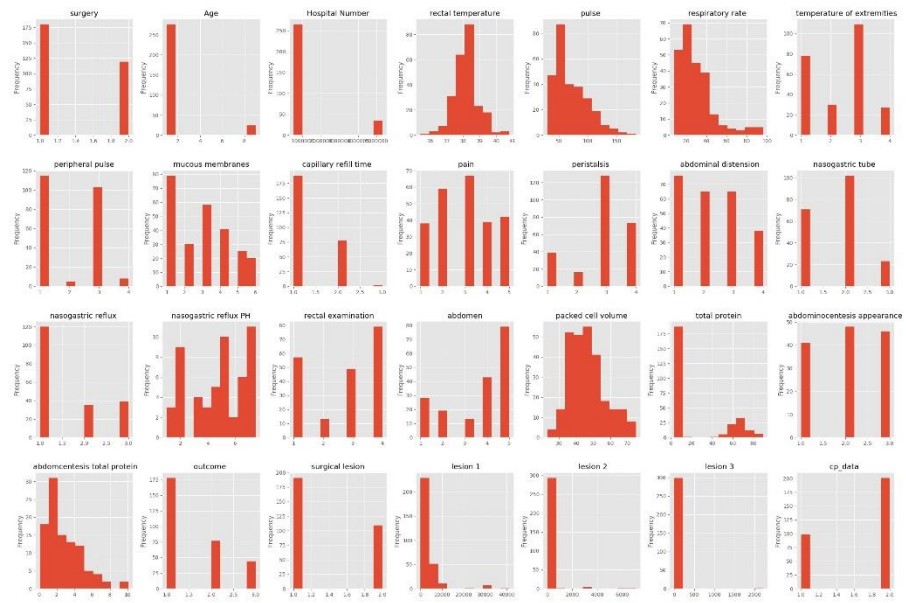
	quartile	missing
surgery	1.00	1.0
Age	1.00	0.0
Hospital Number	528904.00	0.0
rectal temperature	37.80	60.0
pulse	48.00	24.0
respiratory rate	18.50	58.0
temperature of extremities	1.00	56.0
peripheral pulse	1.00	69.0
mucous membranes	1.00	47.0
capillary refill time	1.00	32.0
pain	2.00	55.0
peristalsis	3.00	44.0
abdominal distension	1.00	56.0
nasogastric tube	1.00	104.0
nasogastric reflux	1.00	106.0
nasogastric reflux PH	3.00	247.0
rectal examination	1.00	102.0
abdomen	2.00	118.0
packed cell volume	38.00	29.0
total protein	6.50	33.0
abdominocentesis appearance	1.00	165.0
abdomcentesis total protein	2.00	198.0
outcome	1.00	1.0
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

3.1.2 数据的可视化

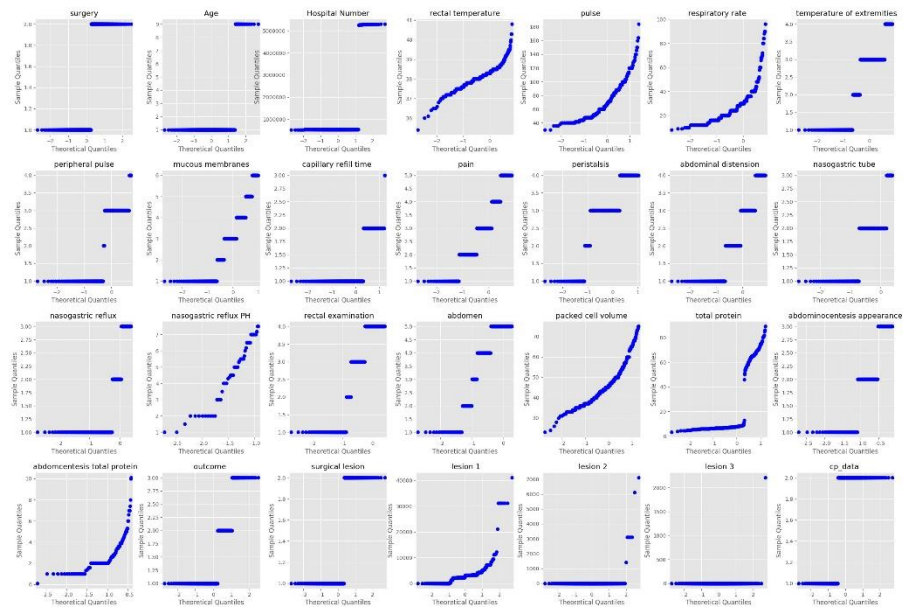
针对数值属性

- 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：

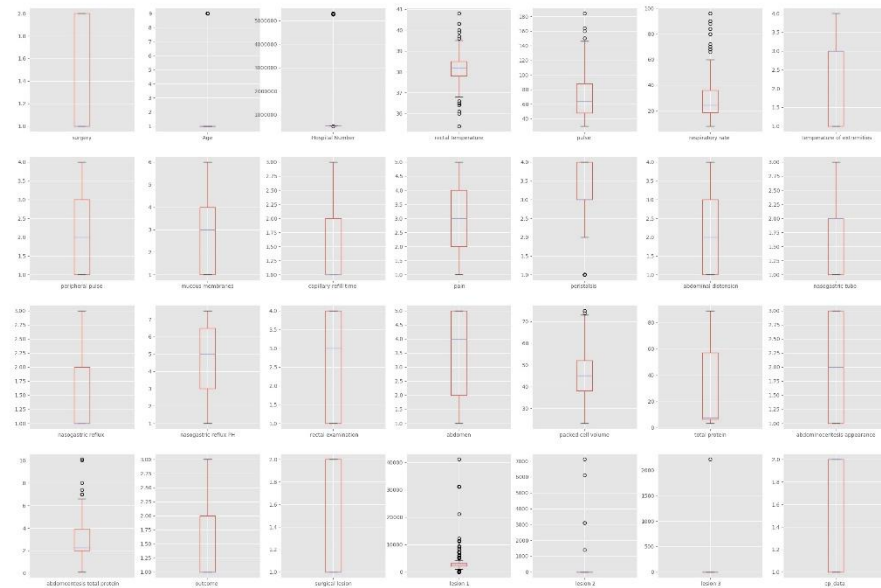


qq 图如下所示:



由各个属性的 qq 图可以看出属性 **rectal temperature** 和属性 **packed cell volume** 满足正态分布

- 绘制盒图，对离群值进行识别



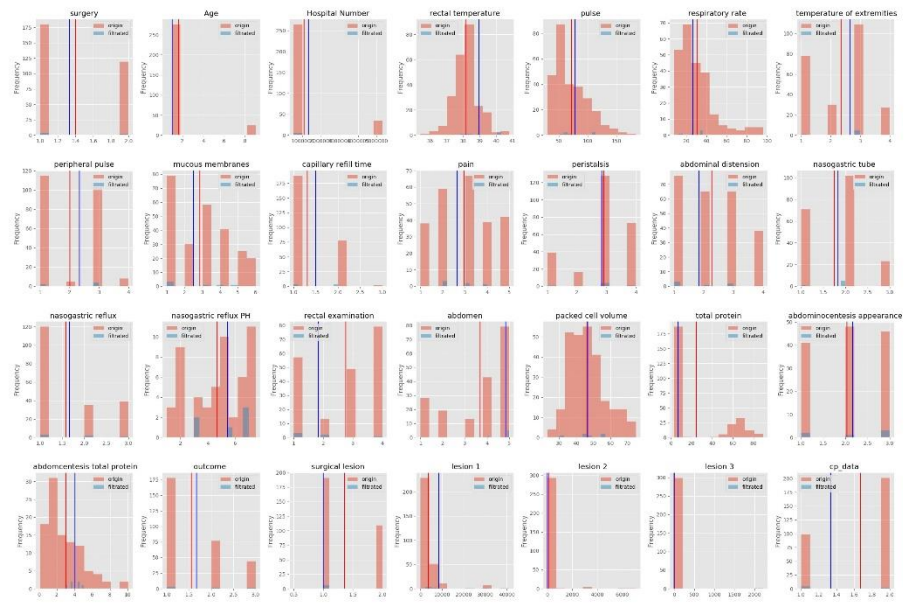
从每个属性的盒图可以看出，属性 Age, Hospital Number, rectal temperature, pulse, respiratory rate, peristalsis, packed cell volume, abdomocentesis total protein, lesion1, lesion2, lesion3 存在离群值

3.2 数据缺失的处理

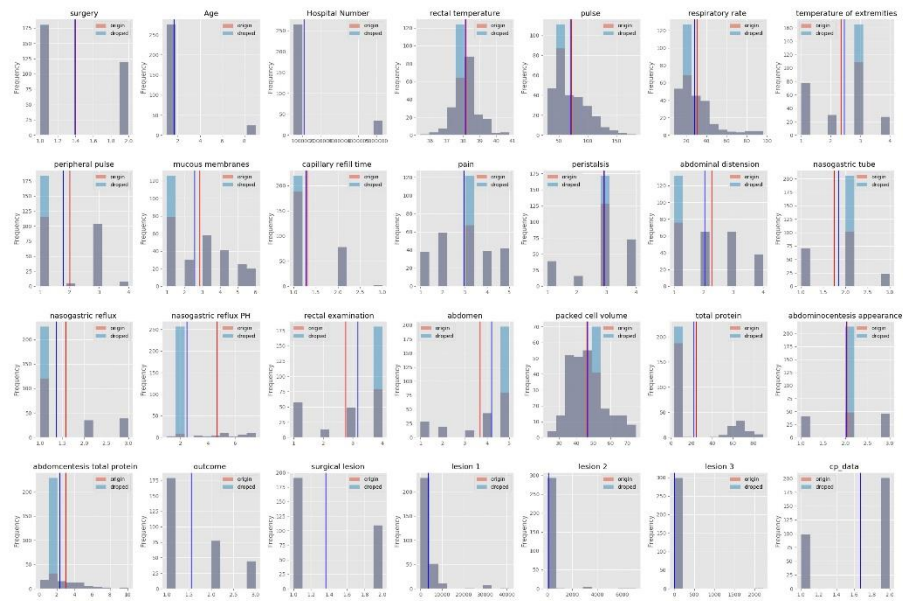
数据集中有 30%的值是缺失的，因此需要先处理数据中的缺失值。

分别使用下列四种策略对缺失值进行处理，处理之后的结果见 data_output 文件夹。

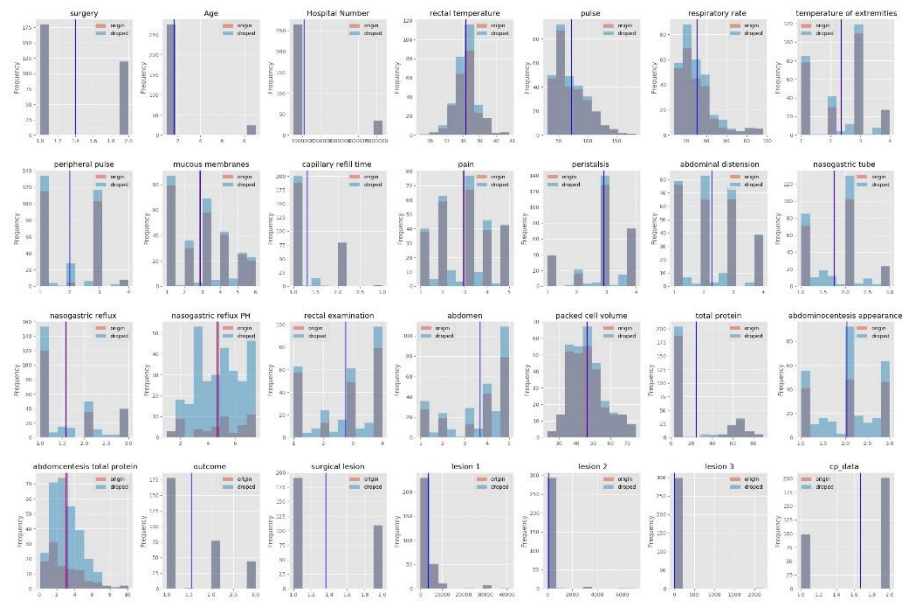
- 将缺失部分剔除



- 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

