马的疝病分析

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: http://archive.ics.uci.edu/ml/datasets/Horse+Colic

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

1. 数据摘要

首先挑选出对称属性和标称属性，分别存入 list 中，标称属性为 name_category，数值属性为 name_value：

```
name_category = ["n1", "n2", "n3", "n7", "n8", "n9", "n10", "n11", "n12", "n13", "n14", "n15", "n17", "n18", "n21", "n23",
name_value = ["v4", "v5", "v6", "v16", "v19", "v20", "v22"]
nameall=["n1", "n2","n3","v4","v5","v6","n7","n8","n9","n10","n11","n12","n13","n14","n15","v16","n17","n18","v19","v20","
```

对标称属性，给出每个可能取值的频数，结果如下：

```
n1
1      214
2      152
dtype: int64
n2
1      340
9       28
dtype: int64
n3
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
n7
3 135
1 95
2 39
4 34
dtype: int64
n8
1 151
3 116
4 12
2 6
dtype: int64
n9
1 98
3 81
4 50
2 38
5 28
6 25
dtype: int64
n10
1 232
2 96
3 2
dtype: int64
n11
3 82
2 77
5 50
1 49
4 47

dtype: int64
n12
3 154
4 91
1 49
2 22
dtype: int64
n13
1 101
3 85
2 75
4 42
dtype: int64
n14
2 121
1 89
3 27
dtype: int64
n15
1 141
3 49
2 45
dtype: int64
n17
4 97
1 68
3 61
2 14
dtype: int64
n18
5 96
4 55
1 31
2 24
3 19
dtype: int64
n21
2 62
3 60
1 52
dtype: int64
n23
1 225
2 89

```
3          52
dtype: int64
n24
1        232
2        136
dtype: int64
n25
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
```

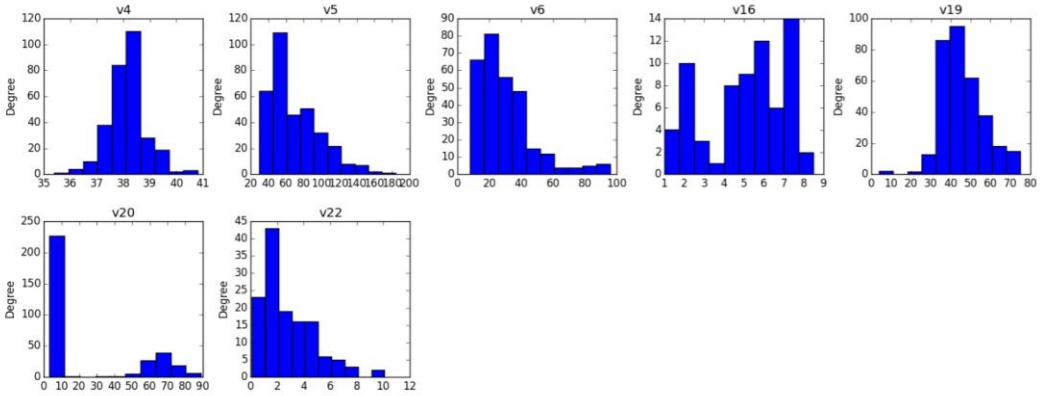
300 1
3115 1
3400 1
41110 1
5000 1
3133 1
12208 1
4122 1
4111 1
5110 1
11300 1
6209 1
9000 1
2305 1
5205 1
7400 1
1111 1
3300 1
8300 1
8405 1
4207 1
21110 1
3207 1
11400 1
dtype: int64
n26
0 358
3111 3
3205 2
6112 1
7111 1
1400 1
2208 1
3112 1
dtype: int64
n27
0 367
2209 1
dtype: int64
n28
2 244
1 124
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：

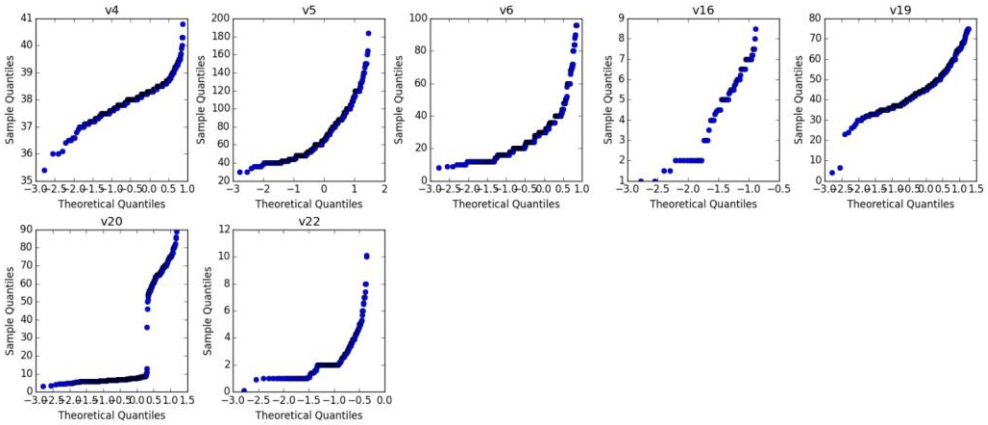
	max	min	mean	median	quartile	missing
v4	40.8	35.4	38.134448	38.1	37.80	69
v5	184.0	30.0	70.757310	60.0	48.00	26
v6	96.0	8.0	30.521886	28.0	18.00	71
v16	8.5	1.0	4.962319	5.4	3.50	299
v19	75.0	4.0	45.656798	44.0	37.25	37
v20	89.0	3.3	24.771077	7.5	6.50	43
v22	10.1	0.1	2.948120	2.1	2.00	235

2. 数据的可视化：

针对数值属性，绘制直方图：

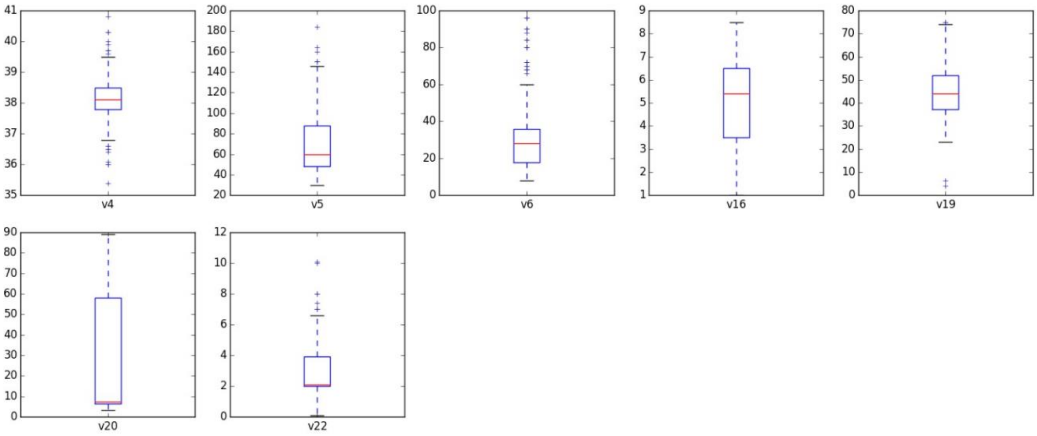


用 qq 图检验其分布是否为正态分布：



可以看出 V4 接近正态分布。

绘制盒图，对离群值进行识别：



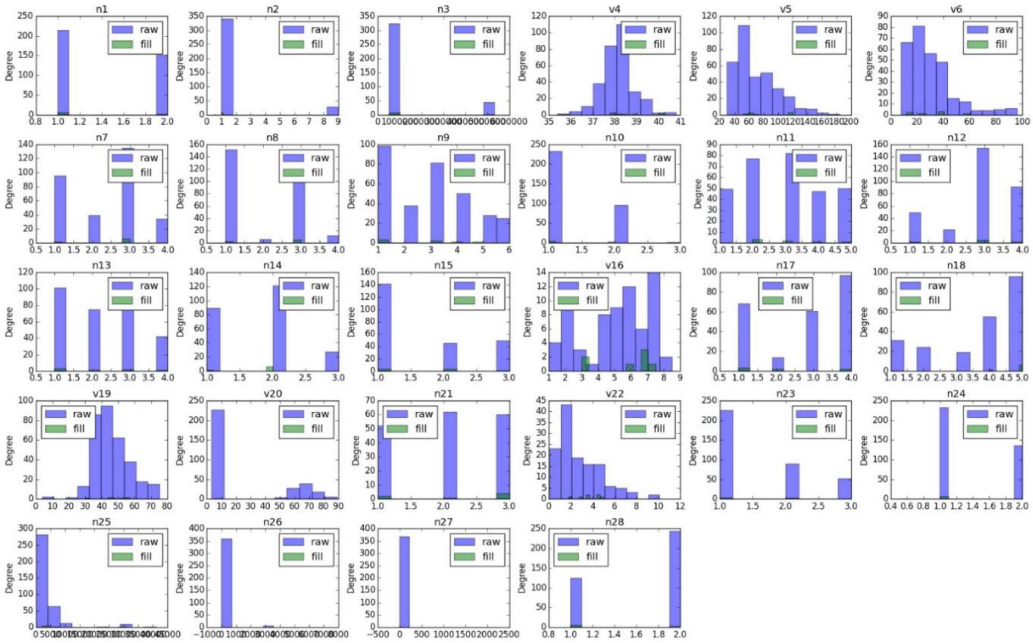
3.2 数据缺失的处理

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:

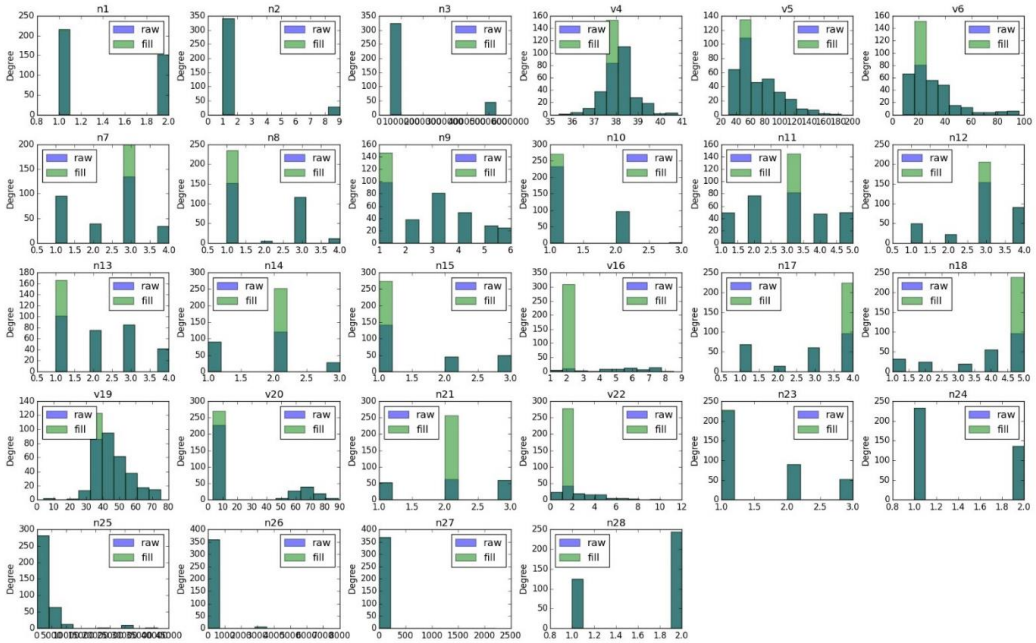
1. 将缺失部分剔除：

   data_fill = data_raw.dropna()，剔除缺失数据。



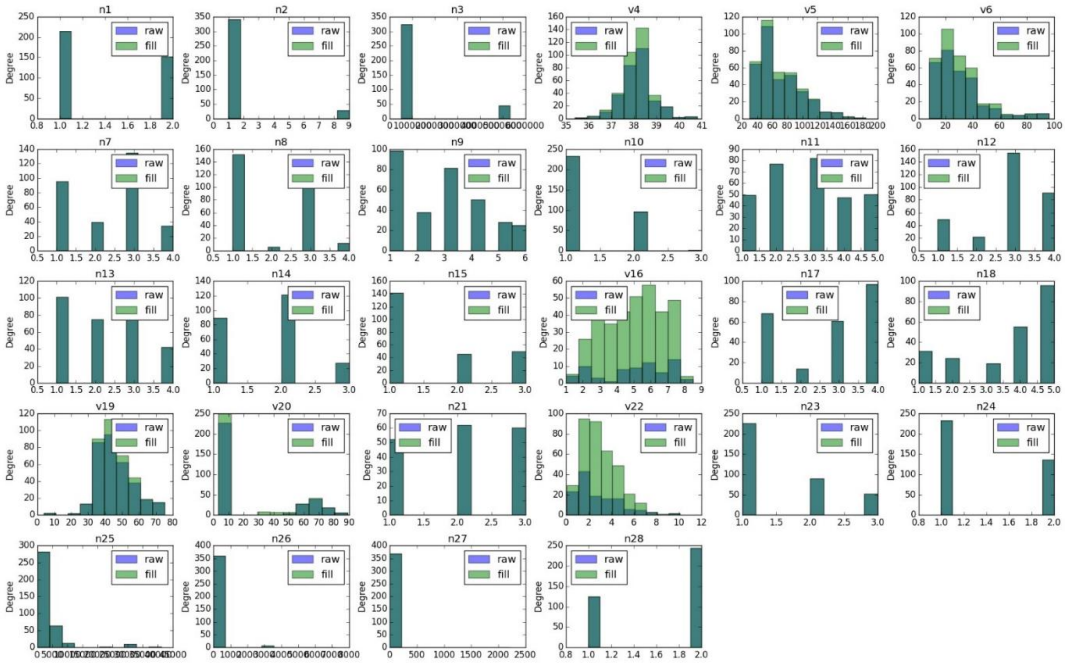
2. 用最高频率值来填补缺失值：

   计算每列最高频率值，填补到空缺位置。



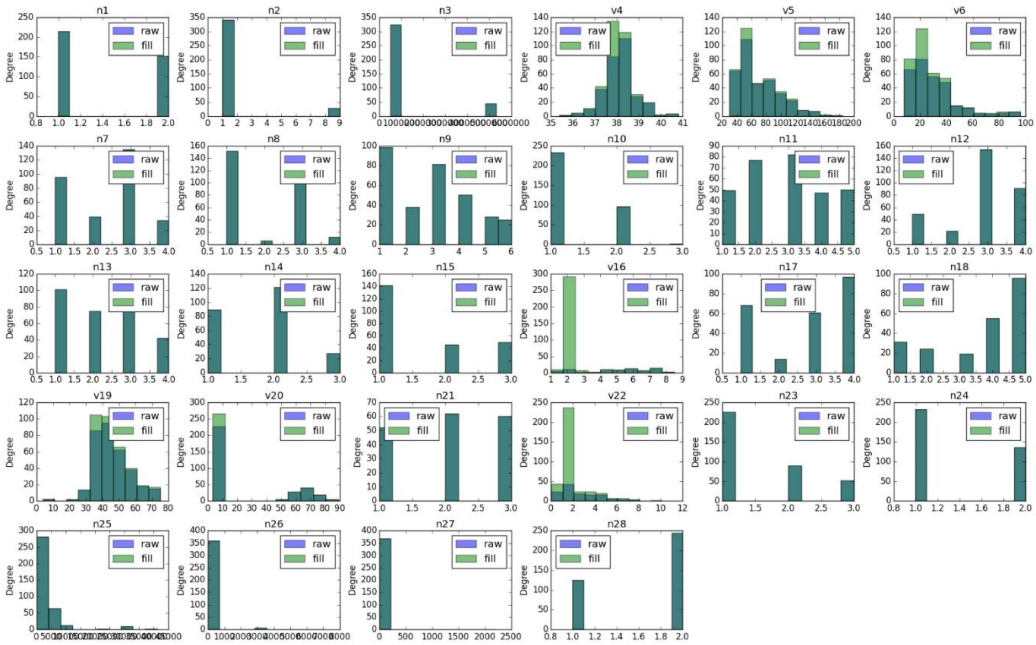
3. 通过属性的相关关系来填补缺失值：

   计算各属性之间相关关系，根据此填补缺失值。



4. 通过数据对象之间的相似性来填补缺失值：

   先正则化，再对各数据对象通过欧氏距离度量其相似性，用相似的数据填补缺失值。


n1
n2
n3
v4
v5
v6
n7
n8
n9
n10
n11
n12
n13
n14
n15
v16
n17
n18
v19
v20
n21
v22
n23
n24
n25
n26
n27
n28
raw
fill
Degree