实验四

崔士强 PB22151743

任务一

Q1

```
Columns with missing values:
['node-caps', 'breast-quad']
```

Q2

```
Value distribution for 'tumor-size' before corrections:
tumor-size
30-34     57
25-29     51
20-24     48
15-19     29
14-Oct    28
40-44     22
35-39     19
0-4        8
50-54      8
9-May      4
45-49      3
Name: count, dtype: int64

Value distribution for 'inv-nodes' before corrections:
inv-nodes
0-2       209
5-Mar      34
8-Jun      17
11-Sep      7
15-17       6
14-Dec      3
24-26       1
Name: count, dtype: int64

Corrected value distribution for 'tumor-size':
tumor-size
30-34    57
25-29    51
20-24    48
15-19    29
10-14    28
40-44    22
35-39    19
0-4       8
50-54     8
5-9       4
45-49     3
Name: count, dtype: int64

Corrected value distribution for 'inv-nodes':
inv-nodes
0-2      209
3-5       34
6-8       17
9-11       7
15-17      6
12-14      3
24-26      1
Name: count, dtype: int64
```

Q3

```
0 Class=no-recurrence-events
1 Class=recurrence-events
2 age=10-19
3 age=20-29
4 age=30-39
5 age=40-49
6 age=50-59
7 age=60-69
8 age=70-79
9 age=80-89
10 age=90-99
11 menopause=lt40
12 menopause=ge40
13 menopause=premeno
14 tumor-size=0-4
15 tumor-size=5-9
16 tumor-size=10-14
17 tumor-size=15-19
18 tumor-size=20-24
19 tumor-size=25-29
20 tumor-size=30-34
21 tumor-size=35-39
22 tumor-size=40-44
23 tumor-size=45-49
24 tumor-size=50-54
25 tumor-size=55-59
26 inv-nodes=0-2
27 inv-nodes=3-5
28 inv-nodes=6-8
29 inv-nodes=9-11
30 inv-nodes=12-14
31 inv-nodes=15-17
32 inv-nodes=18-20
33 inv-nodes=21-23
34 inv-nodes=24-26
35 inv-nodes=27-29
36 inv-nodes=30-32
37 inv-nodes=33-35
38 inv-nodes=36-39
39 node-caps=yes
40 node-caps=no
41 deg-malig=1
42 deg-malig=2
43 deg-malig=3
44 breast=left
45 breast=right
46 breast-quad=left_up
47 breast-quad=left_low
48 breast-quad=right_up
49 breast-quad=right_low
50 breast-quad=central
51 irradiat=yes
52 irradiat=no
```

任务二

Q1

```
频繁项集: [('Class', 0)], 支持度: 0.7075812274368231
频繁项集: [('menopause', 13)], 支持度: 0.5379061371841155
频繁项集: [('menopause', 12)], 支持度: 0.44404332129963897
频繁项集: [('inv-nodes', 26)], 支持度: 0.7545126353790613
频繁项集: [('node-caps', 40)], 支持度: 0.7978339350180506
频繁项集: [('deg-malig', 42)], 支持度: 0.4657039711191336
频繁项集: [('breast', 44)], 支持度: 0.5234657039711191
频繁项集: [('breast', 45)], 支持度: 0.47653429602888087
频繁项集: [('irradiat', 52)], 支持度: 0.776173285198556
频繁项集: [('Class', 0), ('inv-nodes', 26)], 支持度: 0.5992779783393501
频繁项集: [('Class', 0), ('node-caps', 40)], 支持度: 0.6173285198555957
频繁项集: [('Class', 0), ('irradiat', 52)], 支持度: 0.592057761732852
频繁项集: [('inv-nodes', 26), ('menopause', 13)], 支持度: 0.4007220216606498
频繁项集: [('node-caps', 40), ('menopause', 13)], 支持度: 0.4223826714801444
频繁项集: [('irradiat', 52), ('menopause', 13)], 支持度: 0.4007220216606498
频繁项集: [('node-caps', 40), ('inv-nodes', 26)], 支持度: 0.7220216606498195
频繁项集: [('breast', 44), ('inv-nodes', 26)], 支持度: 0.4007220216606498
频繁项集: [('inv-nodes', 26), ('irradiat', 52)], 支持度: 0.6498194945848376
频繁项集: [('breast', 44), ('node-caps', 40)], 支持度: 0.4151624548736462
频繁项集: [('node-caps', 40), ('irradiat', 52)], 支持度: 0.6750902527075813
频繁项集: [('breast', 44), ('irradiat', 52)], 支持度: 0.41155234657039713
频繁项集: [('Class', 0), ('node-caps', 40), ('inv-nodes', 26)], 支持度: 0.5776173285198556
频繁项集: [('Class', 0), ('inv-nodes', 26), ('irradiat', 52)], 支持度: 0.5306859205776173
频繁项集: [('Class', 0), ('node-caps', 40), ('irradiat', 52)], 支持度: 0.5451263537906137
频繁项集: [('Class', 0), ('inv-nodes', 26), ('node-caps', 40)], 支持度: 0.5776173285198556
频繁项集: [('inv-nodes', 26), ('node-caps', 40), ('irradiat', 52)], 支持度: 0.6353790613718412
频繁项集: [('node-caps', 40), ('inv-nodes', 26), ('irradiat', 52)], 支持度: 0.6353790613718412
频繁项集: [('node-caps', 40), ('irradiat', 52), ('Class', 0), ('inv-nodes', 26)], 支持度: 0.5234657039711191
```

Q2

```
规则: [('inv-nodes', 26)] -> ('Class', 0), 置信度: 0.7942583732057417, 提升度: 1.1224978029489308
规则: [('node-caps', 40)] -> ('Class', 0), 置信度: 0.7737556561085973, 提升度: 1.0935220241942933
规则: [('irradiat', 52)] -> ('Class', 0), 置信度: 0.7627906976744185, 提升度: 1.0780256288561936
规则: [('node-caps', 40), ('inv-nodes', 26)] -> ('Class', 0), 置信度: 0.7999999999999999, 提升度: 1.1306122448979592
规则: [('inv-nodes', 26), ('irradiat', 52)] -> ('Class', 0), 置信度: 0.8166666666666667, 提升度: 1.1541666666666668
规则: [('node-caps', 40), ('irradiat', 52)] -> ('Class', 0), 置信度: 0.8074866310160427, 提升度: 1.1411928407726726
规则: [('inv-nodes', 26), ('node-caps', 40)] -> ('Class', 0), 置信度: 0.7999999999999999, 提升度: 1.1306122448979592
规则: [('node-caps', 40), ('irradiat', 52), ('inv-nodes', 26)] -> ('Class', 0), 置信度: 0.8238636363636364, 提升度:
1.1643378942486085
```

Q3

从频繁项集来看，无结节冒，未采取放疗，受侵淋巴结数目在0-2，疾病没有复发这些特征为数据集中较为常见的。

将提取出的关联规则替换为原本的文字得到如下结果：

1. inv-nodes=0-2 -> Class=no-recurrence-events
2. node-caps=no -> Class=no-recurrence-events
3. irradiat=no -> Class=no-recurrence-events
4. node-caps=no, inv-nodes=0-2 -> Class=no-recurrence-events
5. inv-nodes=0-2, irradiat=no -> Class=no-recurrence-events
6. node-caps=no, irradiat=no -> Class=no-recurrence-events
7. inv-nodes=0-2, node-caps=no -> Class=no-recurrence-events
8. node-caps=no, irradiat=no, inv-nodes=0-2 -> Class=no-recurrence-events

由上面的结果可以得到：无结节冒，未采取放疗，受侵淋巴结数目在0-2这些特征存在的情况下，不复发的概率较高。