

# 实验三报告

数据分析及实验 刘沛 PB22061259

## 任务概述

乳腺癌数据集（Breast Cancer Dataset）构建于 1988 年，来源于南斯拉夫卢布尔雅那肿瘤研究所大学医学中心。该数据集记录了 286 个乳腺癌患者的疾病复发情况和部分个体属性值（包含患者年龄、肿瘤大小、是否放疗等 9 种类别型特征）。现欲挖掘该数据集各属性特征之间的频繁项集与关联规则，为乳腺癌的疾病预后提供有用的信息模式，请你按要求编写 Python 代码实现任务列表中的内容。

## 任务列表

1.（25%）读取数据集 **data2.csv**，存储到变量 **df** 中，进行数据预处理。

Q1.（5%）原始数据表存在部分缺失值，请指出哪些特征含有缺失值，并删除所有含空缺值的行。

```
Features with missing values:  
node-caps      8  
breast-quad    1  
dtype: int64
```

```
df = df.dropna()
```

直接一行代码就可以删除具有空缺值的那一行数据

Q2.（10%）当前数据表未能正确处理部分数据值的文本与日期表示类型，使得 **tumor-size** 与 **invnodes** 含有大量异常值，请使用 **value\_counts()**方法验证，并参照 **variables.xlsx** 修正所有异常值。

Value counts for tumor-size:	
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 counts for inv-nodes:	
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	

本来应该是一个范围的值，有的地方却写成了一个日期，的确有这样的错误，对于 14-Oct 这样的值，我们需要对照表格把他修改为 10-14，这样才是正确值，其他的也是以此类推  
修改过后再次检查得到的结果如下：

The screenshot shows a Jupyter Notebook with two output cells. The first cell displays the value counts for 'tumor-size' after correction, and the second cell displays the value counts for 'inv-nodes' after correction. The interface includes a file explorer on the left and a code editor on the right.

26	Value counts for tumor-size:
27	tumor-size
28	30-34 57
29	25-29 51
30	20-24 48
31	15-19 29
32	10-14 28
33	40-44 22
34	35-39 19
35	0-4 8
36	50-54 8
37	5-9 4
38	45-49 3
39	Name: count, dtype: int64
40	
41	Value counts for inv-nodes:
42	inv-nodes
43	0-2 209
44	3-5 34
45	6-8 17
46	9-11 7
47	15-17 6
48	12-14 3
49	24-26 1
50	Name: count, dtype: int64

Q3.（10%）数据表中的特征多为文本属性，不便于后续的关联分析处理过程，请导入 variables.xlsx，用数字索引替换之，并展示索引与属性值的对应关系字典 ind2val。  
例如，Class 属性含 no-recurrence-events 与 recurrence-events 两种可能值，可分别用 0,1 代替，age 含 10-19，20-29 等可能值，可分别用 2,3,...替代之，以此类推。 相应地，可建立字典类型变量：  
ind2val = {0: 'Class=no-recurrence-events', 1: 'Class=recurrence-events', 2: 'age=10-19', 3: 'age=20-29', ... }。

这个问题也很容易解决  
我建立了一个字典，把每一个特征的可能取值都映射为了一个数字，从上到下映射

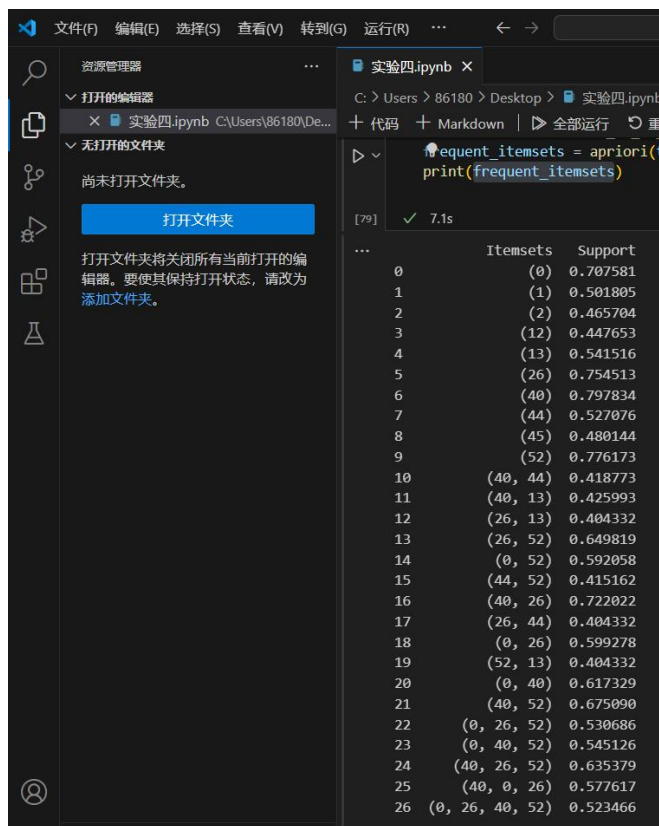
```
{'Class=no-recurrence-events': 0, 'Class=recurrence-events': 1, 'age=10-19': 2, 'age=20-29': 3, 'age=30-39': 4, 'age=40-49': 5, 'age=50-59': 6, 'age=60-69': 7, 'age=70-79': 8, 'age=80-89': 9, 'age=90-99': 10, 'stage=I': 11, 'stage=II': 12, 'stage=III': 13, 'stage=IV': 14, 'stage=V': 15, 'stage=VI': 16, 'stage=VII': 17, 'stage=VIII': 18, 'stage=IX': 19, 'stage=X': 20, 'stage=XI': 21, 'stage=XII': 22, 'stage=XIII': 23, 'stage=XIV': 24, 'stage=XV': 25, 'stage=XVI': 26, 'stage=XVII': 27, 'stage=XVIII': 28, 'stage=XIX': 29, 'stage=XX': 30, 'stage=XXI': 31, 'stage=XXII': 32, 'stage=XXIII': 33, 'stage=XXIV': 34, 'stage=XXV': 35, 'stage=XXVI': 36, 'stage=XXVII': 37, 'stage=XXVIII': 38, 'stage=XXIX': 39, 'stage=XXX': 40, 'stage=XXXI': 41, 'stage=XXXII': 42, 'stage=XXXIII': 43, 'stage=XXXIV': 44, 'stage=XXXV': 45, 'stage=XXXVI': 46, 'stage=XXXVII': 47, 'stage=XXXVIII': 48, 'stage=XXXIX': 49, 'stage=XXXX': 50, 'stage=XXXXI': 51, 'stage=XXXXII': 52, 'stage=XXXXIII': 53, 'stage=XXXXIV': 54, 'stage=XXXXV': 55, 'stage=XXXXVI': 56, 'stage=XXXXVII': 57, 'stage=XXXXVIII': 58, 'stage=XXXXIX': 59, 'stage=XXXXX': 60, 'stage=XXXXXI': 61, 'stage=XXXXXII': 62, 'stage=XXXXXIII': 63, 'stage=XXXXXIV': 64, 'stage=XXXXXV': 65, 'stage=XXXXXVI': 66, 'stage=XXXXXVII': 67, 'stage=XXXXXVIII': 68, 'stage=XXXXXIX': 69, 'stage=XXXXXX': 70, 'stage=XXXXXXI': 71, 'stage=XXXXXXII': 72, 'stage=XXXXXXIII': 73, 'stage=XXXXXXIV': 74, 'stage=XXXXXXV': 75, 'stage=XXXXXXVI': 76, 'stage=XXXXXXVII': 77, 'stage=XXXXXXVIII': 78, 'stage=XXXXXXIX': 79, 'stage=XXXXXXX': 80, 'stage=XXXXXXXI': 81, 'stage=XXXXXXXII': 82, 'stage=XXXXXXXIII': 83, 'stage=XXXXXXXIV': 84, 'stage=XXXXXXXV': 85, 'stage=XXXXXXXVI': 86, 'stage=XXXXXXXVII': 87, 'stage=XXXXXXXVIII': 88, 'stage=XXXXXXXIX': 89, 'stage=XXXXXXXI': 90, 'stage=XXXXXXXII': 91, 'stage=XXXXXXXIII': 92, 'stage=XXXXXXXIV': 93, 'stage=XXXXXXXV': 94, 'stage=XXXXXXXVI': 95, 'stage=XXXXXXXVII': 96, 'stage=XXXXXXXVIII': 97, 'stage=XXXXXXXIX': 98, 'stage=XXXXXXXI': 99, 'stage=XXXXXXXII': 100, 'stage=XXXXXXXIII': 101, 'stage=XXXXXXXIV': 102, 'stage=XXXXXXXV': 103, 'stage=XXXXXXXVI': 104, 'stage=XXXXXXXVII': 105, 'stage=XXXXXXXVIII': 106, 'stage=XXXXXXXIX': 107, 'stage=XXXXXXXI': 108, 'stage=XXXXXXXII': 109, 'stage=XXXXXXXIII': 110, 'stage=XXXXXXXIV': 111, 'stage=XXXXXXXV': 112, 'stage=XXXXXXXVI': 113, 'stage=XXXXXXXVII': 114, 'stage=XXXXXXXVIII': 115, 'stage=XXXXXXXIX': 116, 'stage=XXXXXXXI': 117, 'stage=XXXXXXXII': 118, 'stage=XXXXXXXIII': 119, 'stage=XXXXXXXIV': 120, 'stage=XXXXXXXV': 121, 'stage=XXXXXXXVI': 122, 'stage=XXXXXXXVII': 123, 'stage=XXXXXXXVIII': 124, 'stage=XXXXXXXIX': 125, 'stage=XXXXXXXI': 126, 'stage=XXXXXXXII': 127, 'stage=XXXXXXXIII': 128, 'stage=XXXXXXXIV': 129, 'stage=XXXXXXXV': 130, 'stage=XXXXXXXVI': 131, 'stage=XXXXXXXVII': 132, 'stage=XXXXXXXVIII': 133, 'stage=XXXXXXXIX': 134, 'stage=XXXXXXXI': 135, 'stage=XXXXXXXII': 136, 'stage=XXXXXXXIII': 137, 'stage=XXXXXXXIV': 138, 'stage=XXXXXXXV': 139, 'stage=XXXXXXXVI': 140, 'stage=XXXXXXXVII': 141, 'stage=XXXXXXXVIII': 142, 'stage=XXXXXXXIX': 143, 'stage=XXXXXXXI': 144, 'stage=XXXXXXXII': 145, 'stage=XXXXXXXIII': 146, 'stage=XXXXXXXIV': 147, 'stage=XXXXXXXV': 148, 'stage=XXXXXXXVI': 149, 'stage=XXXXXXXVII': 150, 'stage=XXXXXXXVIII': 151, 'stage=XXXXXXXIX': 152, 'stage=XXXXXXXI': 153, 'stage=XXXXXXXII': 154, 'stage=XXXXXXXIII': 155, 'stage=XXXXXXXIV': 156, 'stage=XXXXXXXV': 157, 'stage=XXXXXXXVI': 158, 'stage=XXXXXXXVII': 159, 'stage=XXXXXXXVIII': 160, 'stage=XXXXXXXIX': 161, 'stage=XXXXXXXI': 162, 'stage=XXXXXXXII': 163, 'stage=XXXXXXXIII': 164, 'stage=XXXXXXXIV': 165, 'stage=XXXXXXXV': 166, 'stage=XXXXXXXVI': 167, 'stage=XXXXXXXVII': 168, 'stage=XXXXXXXVIII': 169, 'stage=XXXXXXXIX': 170, 'stage=XXXXXXXI': 171, 'stage=XXXXXXXII': 172, 'stage=XXXXXXXIII': 173, 'stage=XXXXXXXIV': 174, 'stage=XXXXXXXV': 175, 'stage=XXXXXXXVI': 176, 'stage=XXXXXXXVII': 177, 'stage=XXXXXXXVIII': 178, 'stage=XXXXXXXIX': 179, 'stage=XXXXXXXI': 180, 'stage=XXXXXXXII': 181, 'stage=XXXXXXXIII': 182, 'stage=XXXXXXXIV': 183, 'stage=XXXXXXXV': 184, 'stage=XXXXXXXVI': 185, 'stage=XXXXXXXVII': 186, 'stage=XXXXXXXVIII': 187, 'stage=XXXXXXXIX': 188, 'stage=XXXXXXXI': 189, 'stage=XXXXXXXII': 190, 'stage=XXXXXXXIII': 191, 'stage=XXXXXXXIV': 192, 'stage=XXXXXXXV': 193, 'stage=XXXXXXXVI': 194, 'stage=XXXXXXXVII': 195, 'stage=XXXXXXXVIII': 196, 'stage=XXXXXXXIX': 197, 'stage=XXXXXXXI': 198, 'stage=XXXXXXXII': 199, 'stage=XXXXXXXIII': 200, 'stage=XXXXXXXIV': 201, 'stage=XXXXXXXV': 202, 'stage=XXXXXXXVI': 203, 'stage=XXXXXXXVII': 204, 'stage=XXXXXXXVIII': 205, 'stage=XXXXXXXIX': 206, 'stage=XXXXXXXI': 207, 'stage=XXXXXXXII': 208, 'stage=XXXXXXXIII': 209, 'stage=XXXXXXXIV': 210, 'stage=XXXXXXXV': 211, 'stage=XXXXXXXVI': 212, 'stage=XXXXXXXVII': 213, 'stage=XXXXXXXVIII': 214, 'stage=XXXXXXXIX': 215, 'stage=XXXXXXXI': 216, 'stage=XXXXXXXII': 217, 'stage=XXXXXXXIII': 218, 'stage=XXXXXXXIV': 219, 'stage=XXXXXXXV': 220, 'stage=XXXXXXXVI': 221, 'stage=XXXXXXXVII': 222, 'stage=XXXXXXXVIII': 223, 'stage=XXXXXXXIX': 224, 'stage=XXXXXXXI': 225, 'stage=XXXXXXXII': 226, 'stage=XXXXXXXIII': 227, 'stage=XXXXXXXIV': 228, 'stage=XXXXXXXV': 229, 'stage=XXXXXXXVI': 230, 'stage=XXXXXXXVII': 231, 'stage=XXXXXXXVIII': 232, 'stage=XXXXXXXIX': 233, 'stage=XXXXXXXI': 234, 'stage=XXXXXXXII': 235, 'stage=XXXXXXXIII': 236, 'stage=XXXXXXXIV': 237, 'stage=XXXXXXXV': 238, 'stage=XXXXXXXVI': 239, 'stage=XXXXXXXVII': 240, 'stage=XXXXXXXVIII': 241, 'stage=XXXXXXXIX': 242, 'stage=XXXXXXXI': 243, 'stage=XXXXXXXII': 244, 'stage=XXXXXXXIII': 245, 'stage=XXXXXXXIV': 246, 'stage=XXXXXXXV': 247, 'stage=XXXXXXXVI': 248, 'stage=XXXXXXXVII': 249, 'stage=XXXXXXXVIII': 250, 'stage=XXXXXXXIX': 251, 'stage=XXXXXXXI': 252, 'stage=XXXXXXXII': 253, 'stage=XXXXXXXIII': 254, 'stage=XXXXXXXIV': 255, 'stage=XXXXXXXV': 256, 'stage=XXXXXXXVI': 257, 'stage=XXXXXXXVII': 258, 'stage=XXXXXXXVIII': 259, 'stage=XXXXXXXIX': 260, 'stage=XXXXXXXI': 261, 'stage=XXXXXXXII': 262, 'stage=XXXXXXXIII': 263, 'stage=XXXXXXXIV': 264, 'stage=XXXXXXXV': 265, 'stage=XXXXXXXVI': 266, 'stage=XXXXXXXVII': 267, 'stage=XXXXXXXVIII': 268, 'stage=XXXXXXXIX': 269, 'stage=XXXXXXXI': 270, 'stage=XXXXXXXII': 271, 'stage=XXXXXXXIII': 272, 'stage=XXXXXXXIV': 273, 'stage=XXXXXXXV': 274, 'stage=XXXXXXXVI': 275, 'stage=XXXXXXXVII': 276, 'stage=XXXXXXXVIII': 277, 'stage=XXXXXXXIX': 278, 'stage=XXXXXXXI': 279, 'stage=XXXXXXXII': 280, 'stage=XXXXXXXIII': 281, 'stage=XXXXXXXIV': 282, 'stage=XXXXXXXV': 283, 'stage=XXXXXXXVI': 284, 'stage=XXXXXXXVII': 285, 'stage=XXXXXXXVIII': 286, 'stage=XXXXXXXIX': 287, 'stage=XXXXXXXI': 288, 'stage=XXXXXXXII': 289, 'stage=XXXXXXXIII': 290, 'stage=XXXXXXXIV': 291, 'stage=XXXXXXXV': 292, 'stage=XXXXXXXVI': 293, 'stage=XXXXXXXVII': 294, 'stage=XXXXXXXVIII': 295, 'stage=XXXXXXXIX': 296, 'stage=XXXXXXXI': 297, 'stage=XXXXXXXII': 298, 'stage=XXXXXXXIII': 299, 'stage=XXXXXXXIV': 300, 'stage=XXXXXXXV': 301, 'stage=XXXXXXXVI': 302, 'stage=XXXXXXXVII': 303, 'stage=XXXXXXXVIII': 304, 'stage=XXXXXXXIX': 305, 'stage=XXXXXXXI': 306, 'stage=XXXXXXXII': 307, 'stage=XXXXXXXIII': 308, 'stage=XXXXXXXIV': 309, 'stage=XXXXXXXV': 310, 'stage=XXXXXXXVI': 311, 'stage=XXXXXXXVII': 312, 'stage=XXXXXXXVIII': 313, 'stage=XXXXXXXIX': 314, 'stage=XXXXXXXI': 315, 'stage=XXXXXXXII': 316, 'stage=XXXXXXXIII': 317, 'stage=XXXXXXXIV': 318, 'stage=XXXXXXXV': 319, 'stage=XXXXXXXVI': 320, 'stage=XXXXXXXVII': 321, 'stage=XXXXXXXVIII': 322, 'stage=XXXXXXXIX': 323, 'stage=XXXXXXXI': 324, 'stage=XXXXXXXII': 325, 'stage=XXXXXXXIII': 326, 'stage=XXXXXXXIV': 327, 'stage=XXXXXXXV': 328, 'stage=XXXXXXXVI': 329, 'stage=XXXXXXXVII': 330, 'stage=XXXXXXXVIII': 331, 'stage=XXXXXXXIX': 332, 'stage=XXXXXXXI': 333, 'stage=XXXXXXXII': 334, 'stage=XXXXXXXIII': 335, 'stage=XXXXXXXIV': 336, 'stage=XXXXXXXV': 337, 'stage=XXXXXXXVI': 338, 'stage=XXXXXXXVII': 339, 'stage=XXXXXXXVIII': 340, 'stage=XXXXXXXIX': 341, 'stage=XXXXXXXI': 342, 'stage=XXXXXXXII': 343, 'stage=XXXXXXXIII': 344, 'stage=XXXXXXXIV': 345, 'stage=XXXXXXXV': 346, 'stage=XXXXXXXVI': 347, 'stage=XXXXXXXVII': 348, 'stage=XXXXXXXVIII': 349, 'stage=XXXXXXXIX': 350, 'stage=XXXXXXXI': 351, 'stage=XXXXXXXII': 352, 'stage=XXXXXXXIII': 353, 'stage=XXXXXXXIV': 354, 'stage=XXXXXXXV': 355, 'stage=XXXXXXXVI': 356, 'stage=XXXXXXXVII': 357, 'stage=XXXXXXXVIII': 358, 'stage=XXXXXXXIX': 359, 'stage=XXXXXXXI': 360, 'stage=XXXXXXXII': 361, 'stage=XXXXXXXIII': 362, 'stage=XXXXXXXIV': 363, 'stage=XXXXXXXV': 364, 'stage=XXXXXXXVI': 365, 'stage=XXXXXXXVII': 366, 'stage=XXXXXXXVIII': 367, 'stage=XXXXXXXIX': 368, 'stage=XXXXXXXI': 369, 'stage=XXXXXXXII': 370, 'stage=XXXXXXXIII': 371, 'stage=XXXXXXXIV': 372, 'stage=XXXXXXXV': 373, 'stage=XXXXXXXVI': 374, 'stage=XXXXXXXVII': 375, 'stage=XXXXXXXVIII': 376, 'stage=XXXXXXXIX': 377, 'stage=XXXXXXXI': 378, 'stage=XXXXXXXII': 379, 'stage=XXXXXXXIII': 380, 'stage=XXXXXXXIV': 381, 'stage=XXXXXXXV': 382, 'stage=XXXXXXXVI': 383, 'stage=XXXXXXXVII': 384, 'stage=XXXXXXXVIII': 385, 'stage=XXXXXXXIX': 386, 'stage=XXXXXXXI': 387, 'stage=XXXXXXXII': 388, 'stage=XXXXXXXIII': 389, 'stage=XXXXXXXIV': 390, 'stage=XXXXXXXV': 391, 'stage=XXXXXXXVI': 392, 'stage=XXXXXXXVII': 393, 'stage=XXXXXXXVIII': 394, 'stage=XXXXXXXIX': 395, 'stage=XXXXXXXI': 396, 'stage=XXXXXXXII': 397, 'stage=XXXXXXXIII': 398, 'stage=XXXXXXXIV': 399, 'stage=XXXXXXXV': 400, 'stage=XXXXXXXVI': 401, 'stage=XXXXXXXVII': 402, 'stage=XXXXXXXVIII': 403, 'stage=XXXXXXXIX': 404, 'stage=XXXXXXXI': 405, 'stage=XXXXXXXII': 406, 'stage=XXXXXXXIII': 407, 'stage=XXXXXXXIV': 408, 'stage=XXXXXXXV': 409, 'stage=XXXXXXXVI': 410, 'stage=XXXXXXXVII': 411, 'stage=XXXXXXXVIII': 412, 'stage=XXXXXXXIX': 413, 'stage=XXXXXXXI': 414, 'stage=XXXXXXXII': 415, 'stage=XXXXXXXIII': 416, 'stage=XXXXXXXIV': 417, 'stage=XXXXXXXV': 418, 'stage=XXXXXXXVI': 419, 'stage=XXXXXXXVII': 420, 'stage=XXXXXXXVIII': 421, 'stage=XXXXXXXIX': 422, 'stage=XXXXXXXI': 423, 'stage=XXXXXXXII': 424, 'stage=XXXXXXXIII': 425, 'stage=XXXXXXXIV': 426, 'stage=XXXXXXXV': 427, 'stage=XXXXXXXVI': 428, 'stage=XXXXXXXVII': 429, 'stage=XXXXXXXVIII': 430, 'stage=XXXXXXXIX': 431, 'stage=XXXXXXXI': 432, 'stage=XXXXXXXII': 433, 'stage=XXXXXXXIII': 434, 'stage=XXXXXXXIV': 435, 'stage=XXXXXXXV': 436, 'stage=XXXXXXXVI': 437, 'stage=XXXXXXXVII': 438, 'stage=XXXXXXXVIII': 439, 'stage=XXXXXXXIX': 440, 'stage=XXXXXXXI': 441, 'stage=XXXXXXXII': 442, 'stage=XXXXXXXIII': 443, 'stage=XXXXXXXIV': 444, 'stage=XXXXXXXV': 445, 'stage=XXXXXXXVI': 446, 'stage=XXXXXXXVII': 447, 'stage=XXXXXXXVIII': 448, 'stage=XXXXXXXIX': 449, 'stage=XXXXXXXI': 450, 'stage=XXXXXXXII': 451, 'stage=XXXXXXXIII': 452, 'stage=XXXXXXXIV': 453, 'stage=XXXXXXXV': 454, 'stage=XXXXXXXVI': 455, 'stage=XXXXXXXVII': 456, 'stage=XXXXXXXVIII': 457, 'stage=XXXXXXXIX': 458, 'stage=XXXXXXXI': 459, 'stage=XXXXXXXII': 460, 'stage=XXXXXXXIII': 461, 'stage=XXXXXXXIV': 462, 'stage=XXXXXXXV': 463, 'stage=XXXXXXXVI': 464, 'stage=XXXXXXXVII': 465, 'stage=XXXXXXXVIII': 466, 'stage=XXXXXXXIX': 467, 'stage=XXXXXXXI': 468, 'stage=XXXXXXXII': 469, 'stage=XXXXXXXIII': 470, 'stage=XXXXXXXIV': 471, 'stage=XXXXXXXV': 472, 'stage=XXXXXXXVI': 473, 'stage=XXXXXXXVII': 474, 'stage=XXXXXXXVIII': 475, 'stage=XXXXXXXIX': 476, 'stage=XXXXXXXI': 477, 'stage=XXXXXXXII': 478, 'stage=XXXXXXXIII': 479, 'stage=XXXXXXXIV': 480, 'stage=XXXXXXXV': 481, 'stage=XXXXXXXVI': 482, 'stage=XXXXXXXVII': 483, 'stage=XXXXXXXVIII': 484, 'stage=XXXXXXXIX': 485, 'stage=XXXXXXXI': 486, 'stage=XXXXXXXII': 487, 'stage=XXXXXXXIII': 488, 'stage=XXXXXXXIV': 489, 'stage=XXXXXXXV': 490, 'stage=XXXXXXXVI': 491, 'stage=XXXXXXXVII': 492, 'stage=XXXXXXXVIII': 493, 'stage=XXXXXXXIX': 494, 'stage=XXXXXXXI': 495, 'stage=XXXXXXXII': 496, 'stage=XXXXXXXIII': 497, 'stage=XXXXXXXIV': 498, 'stage=XXXXXXXV': 499, 'stage=XXXXXXXVI': 500, 'stage=XXXXXXXVII': 501, 'stage=XXXXXXXVIII': 502, 'stage=XXXXXXXIX': 503, 'stage=XXXXXXXI': 504, 'stage=XXXXXXXII': 505, 'stage=XXXXXXXIII': 506, 'stage=XXXXXXXIV': 507, 'stage=XXXXXXXV': 508, 'stage=XXXXXXXVI': 509, 'stage=XXXXXXXVII': 510, 'stage=XXXXXXXVIII': 511, 'stage=XXXXXXXIX': 512, 'stage=XXXXXXXI': 513, 'stage=XXXXXXXII': 514, 'stage=XXXXXXXIII': 515, 'stage=XXXXXXXIV': 516, 'stage=XXXXXXXV': 517, 'stage=XXXXXXXVI': 518, 'stage=XXXXXXXVII': 519, 'stage=XXXXXXXVIII': 520, 'stage=XXXXXXXIX': 521, 'stage=XXXXXXXI': 522, 'stage=XXXXXXXII': 523, 'stage=XXXXXXXIII': 524, 'stage=XXXXXXXIV': 525, 'stage=XXXXXXXV': 526, 'stage=XXXXXXXVI': 527, 'stage=XXXXXXXVII': 528, 'stage=XXXXXXXVIII': 529, 'stage=XXXXXXXIX': 530, 'stage=XXXXXXXI': 531, 'stage=XXXXXXXII': 532, 'stage=XXXXXXXIII': 533, 'stage=XXXXXXXIV': 534, 'stage=XXXXXXXV': 535, 'stage=XXXXXXXVI': 536, 'stage=XXXXXXXVII': 537, 'stage=XXXXXXXVIII': 538, 'stage=XXXXXXXIX': 539, 'stage=XXXXXXXI': 540, 'stage=XXXXXXXII': 541, 'stage=XXXXXXXIII': 542, 'stage=XXXXXXXIV': 543, 'stage=XXXXXXXV': 544, 'stage=XXXXXXXVI': 545, 'stage=XXXXXXXVII': 546, 'stage=XXXXXXXVIII': 547, 'stage=XXXXXXXIX': 548, 'stage=XXXXXXXI': 549, 'stage=XXXXXXXII': 550, 'stage=XXXXXXXIII': 551, 'stage=XXXXXXXIV': 552, 'stage=XXXXXXXV': 553, 'stage=XXXXXXXVI': 554, 'stage=XXXXXXXVII': 555, 'stage=XXXXXXXVIII': 556, 'stage=XXXXXXXIX': 557, 'stage=XXXXXXXI': 558, 'stage=XXXXXXXII': 559, 'stage=XXXXXXXIII': 560, 'stage=XXXXXXXIV': 561, 'stage=XXXXXXXV': 562, 'stage=XXXXXXXVI': 563, 'stage=XXXXXXXVII': 564, 'stage=XXXXXXXVIII': 565, 'stage=XXXXXXXIX': 566, 'stage=XXXXXXXI': 567, 'stage=XXXXXXXII': 568, 'stage=XXXXXXXIII': 569, 'stage=XXXXXXXIV': 570, 'stage=XXXXXXXV': 571, 'stage=XXXXXXXVI': 572, 'stage=XXXXXXXVII': 573, 'stage=XXXXXXXVIII': 574, 'stage=XXXXXXXIX': 575, 'stage=XXXXXXXI': 576, 'stage=XXXXXXXII': 577, 'stage=XXXXXXXIII': 578, 'stage=XXXXXXXIV': 579, 'stage=XXXXXXXV': 580, 'stage=XXXXXXXVI': 581, 'stage=XXXXXXXVII': 582, 'stage=XXXXXXXVIII': 583, 'stage=XXXXXXXIX': 584, 'stage=XXXXXXXI': 585, 'stage=XXXXXXXII': 586, 'stage=XXXXXXXIII': 587, 'stage=XXXXXXXIV': 588, 'stage=XXXXXXXV': 589, 'stage=XXXXXXXVI': 590, 'stage=XXXXXXXVII': 591, 'stage=XXXXXXXVIII': 592, 'stage=XXXXXXXIX': 593, 'stage=XXXXXXXI': 594, 'stage=XXXXXXXII': 595, 'stage=XXXXXXXIII': 596, 'stage=XXXXXXXIV': 597, 'stage=XXXXXXXV': 598, 'stage=XXXXXXXVI': 599, 'stage=XXXXXXXVII': 600, 'stage=XXXXXXXVIII': 601, 'stage=XXXXXXXIX': 602, 'stage=XXXXXXXI': 603, 'stage=XXXXXXXII': 604, 'stage=XXXXXXXIII': 605, 'stage=XXXXXXXIV': 606, 'stage=XXXXXXXV': 607, 'stage=XXXXXXXVI': 608, 'stage=XXXXXXXVII': 609, 'stage=XXXXXXXVIII': 610, 'stage=XXXXXXXIX': 611, 'stage=XXXXXXXI': 612, 'stage=XXXXXXXII': 613, 'stage=XXXXXXXIII': 614, 'stage=XXXXXXXIV': 615, 'stage=XXXXXXXV': 616, 'stage=XXXXXXXVI': 617, 'stage=XXXXXXXVII': 618, 'stage=XXXXXXXVIII': 619, 'stage=XXXXXXXIX': 620, 'stage=XXXXXXXI': 621, 'stage=XXXXXXXII': 622, 'stage=XXXXXXXIII': 623, 'stage=XXXXXXXIV': 624, 'stage=XXXXXXXV': 625, 'stage=XXXXXXXVI': 626, 'stage=XXXXXXXVII': 627, 'stage=XXXXXXXVIII': 628, 'stage=XXXXXXXIX': 629, 'stage=XXXXXXXI': 630, 'stage=XXXXXXXII': 631, 'stage=XXXXXXXIII': 632, 'stage=XXXXXXXIV': 633, 'stage=XXXXXXXV': 634, 'stage=XXXXXXXVI': 635, 'stage=XXXXXXXVII': 636, 'stage=XXXXXXXVIII': 637, 'stage=XXXXXXXIX': 638, 'stage=XXXXXXXI': 639, 'stage=XXXXXXXII': 640, 'stage=XXXXXXXIII': 641, 'stage=XXXXXXXIV': 642, 'stage=XXXXXXXV': 643, 'stage=XXXXXXXVI': 644, 'stage=XXXXXXXVII': 645, 'stage=XXXXXXXVIII': 646, 'stage=XXXXXXXIX': 647, 'stage=XXXXXXXI': 648, 'stage=XXXXXXXII': 649, 'stage=XXXXXXXIII': 650, 'stage=XXXXXXXIV': 651, 'stage=XXXXXXXV': 652, 'stage=XXXXXXXVI': 653, 'stage=XXXXXXXVII': 654, 'stage=XXXXXXXVIII': 655, 'stage=XXXXXXXIX': 656, 'stage=XXXXXXXI': 657, 'stage=XXXXXXXII': 658, 'stage=XXXXXXXIII': 659, 'stage=XXXXXXXIV': 660, 'stage=XXXXXXXV': 661, 'stage=XXXXXXXVI': 662, 'stage=XXXXXXXVII': 663, 'stage=XXXXXXXVIII': 664, 'stage=XXXXXXXIX': 665, 'stage=XXXXXXXI': 666, 'stage=XXXXXXXII': 667, 'stage=XXXXXXXIII': 668, 'stage=XXXXXXXIV': 669, 'stage=XXXXXXXV': 670, 'stage=XXXXXXXVI': 671, 'stage=XXXXXXXVII': 672, 'stage=XXXXXXXVIII': 673, 'stage=XXXXXXXIX': 674, 'stage=XXXXXXXI': 675, 'stage=XXXXXXXII': 676, 'stage=XXXXXXXIII': 677, 'stage=XXXXXXXIV': 678, 'stage=XXXXXXXV': 679, 'stage=XXXXXXXVI': 680, 'stage=XXXXXXXVII': 681, 'stage=XXXXXXXVIII': 682, 'stage=XXXXXXXIX': 683, 'stage=XXXXXXXI': 684, 'stage=XXXXXXXII': 685, 'stage=XXXXXXXIII': 686, 'stage=XXXXXXXIV': 687, 'stage=XXXXXXXV': 688, 'stage=XXXXXXXVI': 689, 'stage=XXXXXXXVII': 690, 'stage=XXXXXXXVIII': 691, 'stage=XXXXXXXIX': 692, 'stage=XXXXXXXI': 693, 'stage=XXXXXXXII': 694, 'stage=XXXXXXXIII': 695, 'stage=XXXXXXXIV': 696, 'stage=XXXXXXXV': 697, 'stage=XXXXXXXVI': 698, 'stage=XXXXXXXVII': 699, 'stage=XXXXXXXVIII': 700, 'stage=XXXXXXXIX': 701, 'stage=XXXXXXXI': 702, 'stage=XXXXXXXII': 703, 'stage=XXXXXXXIII': 704, 'stage=XXXXXXXIV': 705, 'stage=XXXXXXXV': 706, 'stage=XXXXXXXVI': 707, 'stage=XXXXXXXVII': 708, 'stage=XXXXXXXVIII': 709, 'stage=XXXXXXXIX': 710, 'stage=XXXXXXXI': 711, 'stage=XXXXXXXII': 712, 'stage=XXXXXXXIII': 713, 'stage=XXXXXXXIV': 714, 'stage=XXXXXXXV': 715, 'stage=XXXXXXXVI': 716, 'stage=XXXXXXXVII': 717, 'stage=XXXXXXXVIII': 718, 'stage=XXXXXXXIX': 719, 'stage=XXXXXXXI': 720, 'stage=XXXXXXXII': 721, 'stage=XXXXXXXIII': 722, 'stage=XXXXXXXIV': 723, 'stage=XXXXXXXV': 724, 'stage=XXXXXXXVI': 725, 'stage=XXXXXXXVII': 726, 'stage=XXXXXXXVIII': 727, 'stage=XXXXXXXIX': 728, 'stage=XXXXXXXI': 729, 'stage=XXXXXXXII': 730, 'stage=XXXXXXXIII': 731, 'stage=XXXXXXXIV': 732, 'stage=XXXXXXXV': 733, 'stage=XXXXXXXVI': 734, 'stage=XXXXXXXVII': 735, 'stage=XXXXXXXVIII': 736, 'stage=XXXXXXXIX': 737, 'stage=XXXXXXXI': 738, 'stage=XXXXXXXII': 739, 'stage=XXXXXXXIII': 740, 'stage=XXXXXXXIV': 741, 'stage=XXXXXXXV': 742, 'stage=XXXXXXXVI': 743, 'stage=XXXXXXXVII': 744, 'stage=XXXXXXXVIII': 745, 'stage=XXXXXXXIX': 746, 'stage=XXXXXXXI': 747, 'stage=XXXXXXXII': 748, 'stage=XXXXXXXIII': 749, 'stage=XXXXXXXIV': 750, 'stage=XXXXXXXV': 751, 'stage=XXXXXXXVI': 752, 'stage=XXXXXXXVII': 753, 'stage=XXXXXXXVIII': 754, 'stage=XXXXXXXIX': 755, 'stage=XXXXXXXI': 756, 'stage=XXXXXXXII': 757, 'stage=XXXXXXXIII': 758, 'stage=XXXXXXXIV': 759, 'stage=XXXXXXXV': 760, 'stage=XXXXXXXVI': 761, 'stage=XXXXXXXVII': 762, 'stage=XXXXXXXVIII': 763, 'stage=XXXXXXXIX': 764, 'stage=XXXXXXXI': 765, 'stage=XXXXXXXII': 766, 'stage=XXXXXXXIII': 767, 'stage=XXXXXXXIV': 768, 'stage=XXXXXXXV': 769, 'stage=XXXXXXXVI': 770, 'stage=XXXXXXXVII': 771, 'stage=XXXXXXXVIII': 772, 'stage=XXXXXXXIX': 773, 'stage=XXXXXXXI': 774, 'stage=XXXXXXXII': 775, 'stage=XXXXXXXIII': 776, 'stage=XXXXXXXIV': 777, 'stage=XXXXXXXV': 778, 'stage=XXXXXXXVI': 779, 'stage=XXXXXXXVII': 780, 'stage=XXXXXXXVIII': 781, 'stage=XXXXXXXIX': 782, 'stage=XXXXXXXI': 783, 'stage=XXXXXXXII': 784, 'stage=XXXXXXXIII': 785, 'stage=XXXXXXXIV': 786, 'stage=XXXXXXXV': 787, 'stage=XXXXXXXVI': 788, 'stage=XXXXXXXVII': 789, 'stage=XXXXXXXVIII': 790, 'stage=XXXXXXXIX': 791, 'stage=XXXXXXXI': 792, 'stage=XXXXXXXII': 793, 'stage=XXXXXXXIII': 794, 'stage=XXXXXXXIV': 795, 'stage=XXXXXXXV': 796, 'stage=XXXXXXXVI': 797, 'stage=XXXXXXXVII': 798, 'stage=XXXXXXXVIII': 799, 'stage=XXXXXXXIX': 800, 'stage=XXXXXXXI': 801, 'stage=XXXXXXXII': 802, 'stage=XXXXXXXIII': 803, 'stage=XXXXXXXIV': 804, 'stage=XXXXXXXV': 805, 'stage=XXXXXXXVI': 806, 'stage=XXXXXXXVII': 807, 'stage=XXXXXXXVIII': 808, 'stage=XXXXXXXIX': 809, 'stage=XXXXXXXI': 810, 'stage=XXXXXXXII': 811, 'stage=XXXXXXXIII': 812, 'stage=XXXXXXXIV': 813, 'stage=XXXXXXXV': 814, 'stage=XXXXXXXVI': 815, 'stage=XXXXXXXVII': 816, 'stage=XXXXXXXVIII': 817, 'stage=XXXXXXXIX': 818, 'stage=XXXXXXXI': 819, 'stage=XXXXXXXII': 820, 'stage=XXXXXXXIII': 821, 'stage=XXXXXXXIV': 822, 'stage=XXXXXXXV': 823, 'stage=XXXXXXXVI': 824, 'stage=XXXXXXXVII': 825, 'stage=XXXXXXXVIII': 826, 'stage=XXXXXXXIX': 827, 'stage=XXXXXXXI': 828, 'stage=XXXXXXXII': 829, 'stage=XXXXXXXIII': 830, 'stage=XXXXXXXIV': 831, 'stage=XXXXXXXV': 832, 'stage=XXXXXXXVI': 833, 'stage=XXXXXXXVII': 834, 'stage=XXXXXXXVIII': 835, 'stage=XXXXXXXIX': 836, 'stage=XXXXXXXI': 837, 'stage=XXXXXXXII': 838, 'stage=XXXXXXXIII': 839, 'stage=XXXXXXXIV': 840, 'stage=XXXXXXXV': 841, 'stage=XXXXXXXVI': 842, 'stage=XXXXXXXVII': 843, 'stage=XXXXXXXVIII': 844, 'stage=XXXXXXXIX': 845, 'stage=XXXXXXXI': 846, 'stage=XXXXXXXII': 847, 'stage=XXXXXXXIII': 848, 'stage=XXXXXXXIV': 849, 'stage=XXXXXXXV': 850, 'stage=XXXXXXXVI': 851, 'stage=XXXXXXXVII': 852, 'stage=XXXXXXXVIII': 853, 'stage=XXXXXXXIX': 854, 'stage=XXXXXXXI': 855, 'stage=XXXXXXXII': 856, 'stage=XXXXXXXIII': 857, 'stage=XXXXXXXIV': 858, 'stage=XXXXXXXV': 859, 'stage=XXXXXXXVI': 860, 'stage=XXXXXXXVII': 861, 'stage=XXXXXXXVIII': 862, 'stage=XXXXXXXIX': 863, 'stage=XXXXXXXI': 864, 'stage=XXXXXXXII': 865, 'stage=XXXXXXXIII': 866, 'stage=XXXXXXXIV': 867, 'stage=XXXXXXXV': 868, 'stage=XXXXXXXVI': 869, 'stage=XXXXXXXVII': 870, 'stage=XXXXXXXVIII': 871, 'stage=XXXXXXXIX': 872, 'stage=XXXXXXXI': 873, 'stage=XXXXXXXII': 874, 'stage=XXXXXXXIII': 875, 'stage=XXXXXXXIV': 876, 'stage=XXXXXXXV': 877, 'stage=XXXXXXXVI': 878, 'stage=XXXXXXXVII': 879, 'stage=XXXXXXXVIII': 880, 'stage=XXXXXXXIX': 881, 'stage=XXXXXXXI': 882, 'stage=XXXXXXXII': 883, 'stage=XXXXXXXIII': 884, 'stage=XXXXXXXIV': 885, 'stage=XXXXXXXV': 886, 'stage=XXXXXXXVI': 887, 'stage=XXXXXXXVII': 888, 'stage=XXXXXXXVIII': 889, 'stage=XXXXXXXIX': 890, 'stage=XXXXXXXI': 891, 'stage=XXXXXXXII': 892, 'stage=XXXXXXXIII': 893, 'stage=XXXXXXXIV': 894, 'stage=XXXXXXXV': 895, 'stage=XXXXXXXVI': 896, 'stage=XXXXXXXVII': 897, 'stage=XXXXXXXVIII': 898, 'stage=XXXXXXXIX': 899, 'stage=XXXXXXXI': 900, 'stage=XXXXXXXII': 901, 'stage=XXXXXXXIII': 902, 'stage=XXXXXXXIV': 903, 'stage=XXXXXXXV': 904, 'stage
```

```
'breast-quad=right_up': 48, 'breast-quad=right_low': 49, 'breast-quad=central': 50, 'irradiat=yes': 51, 'irradiat=no': 52}
```

## 2.（75%）基于预处理后的数据集 df，编写算法代码进行关联规则分析。

Q1.（45%）请参考以下 Apriori 产生频繁项集的算法流程，自行编写相应代码，以最小支持度阈值为 0.4，挖掘 df 中的频繁项集。

产生的频繁项集如下图所示：



	Itemsets	Support
0	(0)	0.707581
1	(1)	0.501805
2	(2)	0.465704
3	(12)	0.447653
4	(13)	0.541516
5	(26)	0.754513
6	(40)	0.797834
7	(44)	0.527076
8	(45)	0.480144
9	(52)	0.776173
10	(40, 44)	0.418773
11	(40, 13)	0.425993
12	(26, 13)	0.404332
13	(26, 52)	0.649819
14	(0, 52)	0.592058
15	(44, 52)	0.415162
16	(40, 26)	0.722022
17	(26, 44)	0.404332
18	(0, 26)	0.599278
19	(52, 13)	0.404332
20	(0, 40)	0.617329
21	(40, 52)	0.675090
22	(0, 26, 52)	0.530686
23	(0, 40, 52)	0.545126
24	(40, 26, 52)	0.635379
25	(40, 0, 26)	0.577617
26	(0, 26, 40, 52)	0.523466

Q2.（20%）基于提取出的频繁项集，以最小置信度阈值为 0.75，提取形如  $X \rightarrow \{0\}$  的强关联规则，并分别输出它们的置信度和提升度。

置信度的计算方法： $\text{Confidence}(X \rightarrow Y) = \text{support}(X, Y) / \text{support}(X)$

这里我们只考虑后件即  $X$  为  $\{0\}$  的情况，而前件没有任何限定。

提升度： $\text{Lift}(X \rightarrow Y) = \text{Confidence}(X \rightarrow Y) / \text{Support}(Y)$

输出的结果如下图所示：

其中每一条关联规则是一个元组，第一项为前件，第二项为后件。第三项是置信度，第四项是提升度

```
({52}, 0, 0.7627906976744185, 1.0780256288561936)
({26}, 0, 0.7942583732057417, 1.1224978029489308)
({40}, 0, 0.7737556561085973, 1.0935220241942933)
({26, 52}, 0, 0.8166666666666667, 1.1541666666666668)
({40, 52}, 0, 0.8074866310160427, 1.1411928407726726)
({40, 26}, 0, 0.7999999999999999, 1.1306122448979592)
({40, 26, 52}, 0, 0.8238636363636364, 1.1643378942486085)
```

可以看得出来，每一项的提升度都大于 1

#2.Q3. (10%) 参考 ind2val 中索引与属性值的对应关系，对以上频繁项集和关联规则结果进行简要分析和总结。

把数字索引逆映射回去，得到的关联规则如下图所示

```
({'irradiat=no'}, 'Class=no-recurrence-events', 0.7627906976744185, 1.0780256288561936)
({'inv-nodes=0-2'}, 'Class=no-recurrence-events', 0.7942583732057417, 1.1224978029489308)
({'node-caps=no'}, 'Class=no-recurrence-events', 0.7737556561085973, 1.0935220241942933)
({'irradiat=no', 'inv-nodes=0-2'}, 'Class=no-recurrence-events', 0.8166666666666667, 1.1541666666666668)
({'irradiat=no', 'node-caps=no'}, 'Class=no-recurrence-events', 0.8074866310160427, 1.1411928407726726)
({'inv-nodes=0-2', 'node-caps=no'}, 'Class=no-recurrence-events', 0.7999999999999999, 1.1306122448979592)
({'irradiat=no', 'inv-nodes=0-2', 'node-caps=no'}, 'Class=no-recurrence-events', 0.8238636363636364, 1.1643378942486085)
```

可以发现，其实在这样的阈值设置下，得到的可行关联规则也还是错综繁杂，难以分析，不过我们还是大略可以根据得到的这些规则，做出以下定论：

1. 没有放疗经历的乳腺癌患者一般不会复发
2. 受侵淋巴结数目非常少的乳腺癌患者一般不会复发
3. 结冒节数目非常少的乳腺癌患者一般不会复发
4. ....

最终结论：

放疗经历、受侵淋巴结数目大小、结冒节数目大小这三项参数几乎可以很好的描述出乳腺癌患者是否会复发

而没有放疗经历的、受侵淋巴结数目非常少的、几乎没有结冒节的乳腺癌患者一般不会复发

其实，我觉得这样的规则寻找还是应该把前件和后件分为两个集合。就比如说，这次实验当中我们希望知道哪些因素对乳腺癌患者是否复发的影响比较大，于是就把乳腺癌患者不复发作为后件去寻找相应的强关联规则。

考虑到数据的现实意义，后件应该是某种抽象的结论，前件应该是某种症状，这样的关联规则可以帮助医生进行诊断，总结出得某种病会有哪些症状。或者，后件是得了某种病，前件是病人的其他一些特征（包括病史和生活习惯等），也可以总结出该病的高发人群

所以，关联规则的得到，其实还需要考虑主观目的，既我要用这个数据集来得到一些什么样

的结论，否则，把所有特征都打乱在一起胡乱分析，得到的关联规则也是特别特别多，很多规则意义还不明确。

不过话说回来，我们作为数据的学生其实并不太懂医学相关的东西，做数据分析的时候自然也是苍蝇乱撞，只是为了分析而分析罢了。我们实际上也不明白，淋巴结是什么，数目少背后的含义是什么，我甚至都没有听说过结直肠，更不用说去分析其数目背后代表的医学含义了！

所以，未来真正的要去做数据分析工作的时候，肯定要事先学习了解一些数据集专业的相关知识，或者与相关专业领域的专家进行合作吧！