In [1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as mp
import numpy as np

hm=pd.read_csv("haberman.csv")
```

In [2]:
```
print(hm.shape)
'''number of points and features of the data set are'''
```

```
(306, 4)
```

In [4]:
```
print(hm.columns)
'''number of columns of data set are'''
```

```
Index(['age', 'year', 'nodes', 'status'], dtype='object')
```

In [5]:
```
hm["age"].value_counts()
```

Out[5]:
```
52    14
54    13
50    12
47    11
53    11
43    11
57    11
55    10
65    10
49    10
38    10
41    10
61     9
45     9
42     9
63     8
59     8
```

```
62     7
44     7
58     7
56     7
46     7
70     7
34     7
48     7
37     6
67     6
60     6
51     6
39     6
66     5
64     5
72     4
69     4
40     3
30     3
68     2
73     2
74     2
36     2
35     2
33     2
31     2
78     1
71     1
75     1
76     1
77     1
83     1
Name: age, dtype: int64
```

In [6]:
```
hm["year"].value_counts()
```

Out[6]:
```
58    36
64    31
63    30
```

```
66    28
65    28
60    28
59    27
61    26
67    25
62    23
68    13
69    11
Name: year, dtype: int64
```

In [7]:
```
hm["nodes"].value_counts()
```

Out[7]:
```
0     136
1      41
2      20
3      20
4      13
6       7
7       7
8       7
5       6
9       6
13      5
14      4
11      4
10      3
15      3
19      3
22      3
23      3
12      2
20      2
46      1
16      1
17      1
18      1
21      1
24      1
```

```
25      1
28      1
30      1
35      1
52      1
Name: nodes, dtype: int64
```

In [8]:
```
hm["status"].value_counts()
''' haberman data set is an imbalanced data set with classes 1 and 2'''
```

Out[8]:
```
1    225
2     81
Name: status, dtype: int64
```

In [1]:
```
'''by the given data it is found that the data set contains columns nam
ed age, year, nodes and status.
where the status represents weather the patient lives for more than 5 y
ears or less than 5years,
if the survival status is 1 then the patient may survive for 5 years or
 longer if the status is 2 then the patient is
dead within 5 years

here our main objective is to find wheather the patient will survive fo
r more than 5 years or not based upon the patients age,
year of treatment and the number of lymph nodes.'''


'''pdf'''


sns.FacetGrid(hm, hue="status", size=5) \
   .map(sns.distplot, "age") \
   .add_legend();
plt.show();
```

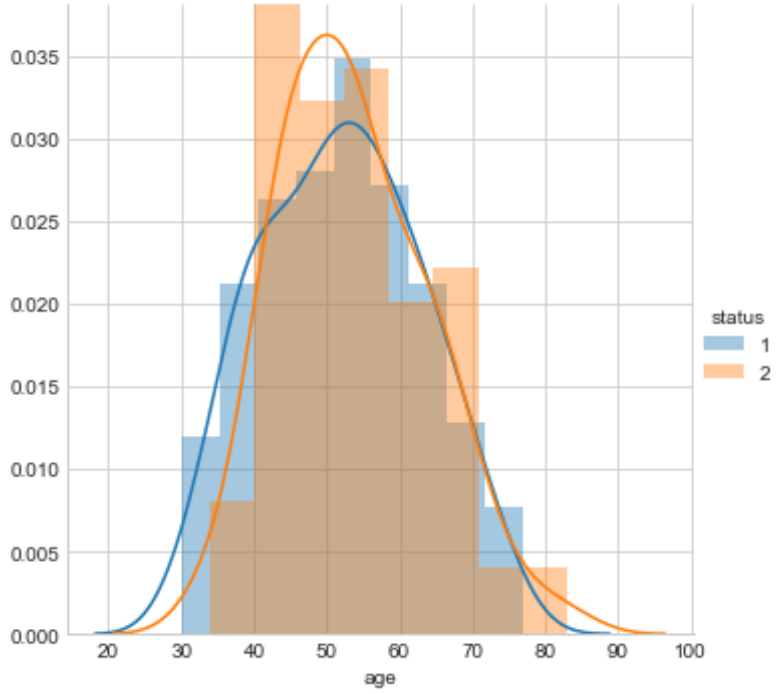
```
---------------------------------------------------------------------
----
NameError                                 Traceback (most recent call l
ast)
```

```
<ipython-input-1-a5ea5d0e1925> in <module>()
     11
     12
---> 13 sns.FacetGrid(hm, hue="status", size=5)    .map(sns.distplot,
"age")    .add_legend();
     14 plt.show();

NameError: name 'sns' is not defined
```

In [2]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

hm=pd.read_csv("haberman.csv")
```

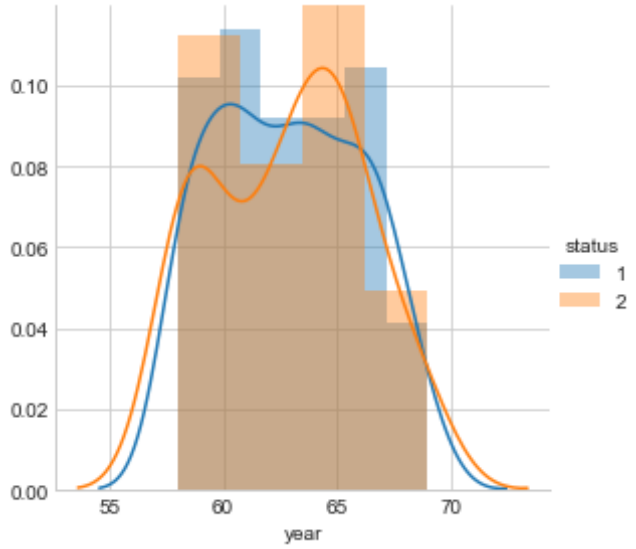
In [5]:
```
sns.FacetGrid(hm, hue="status", size=5) \
   .map(sns.distplot,"age") \
   .add_legend();
plt.show();
```

```
C:\Users\chandu1\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6
462: UserWarning: The 'normed' kwarg is deprecated, and has been replac
ed by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\chandu1\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6
462: UserWarning: The 'normed' kwarg is deprecated, and has been replac
ed by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



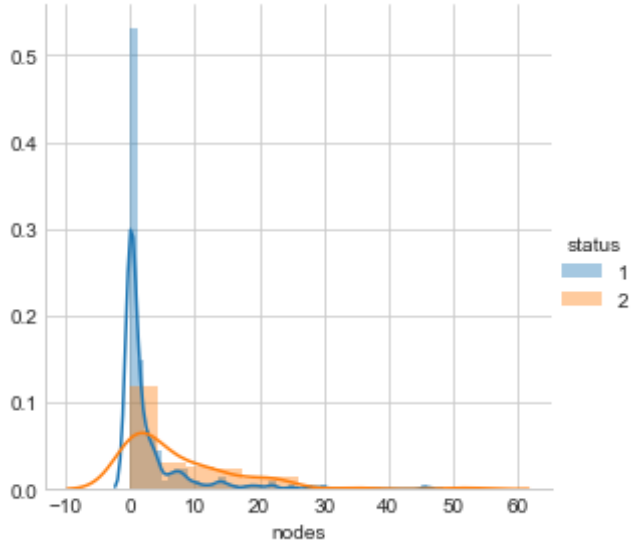
In [8]:
```
sns.FacetGrid(hm, hue="status", size=4) \
   .map(sns.distplot, "year") \
   .add_legend();
plt.show();
```

```
C:\Users\chandu1\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6
462: UserWarning: The 'normed' kwarg is deprecated, and has been replac
ed by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\chandu1\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6
462: UserWarning: The 'normed' kwarg is deprecated, and has been replac
ed by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



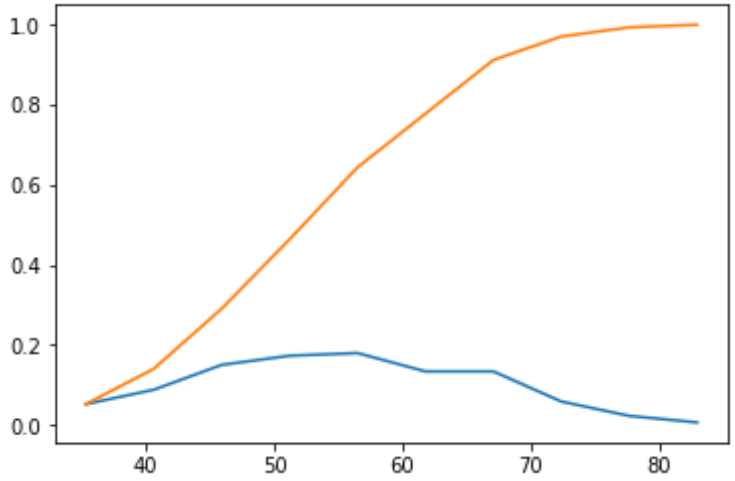
In [9]:
```
sns.FacetGrid(hm, hue="status", size=4) \
   .map(sns.distplot, "nodes") \
   .add_legend();
plt.show();
```

```
C:\Users\chandu1\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6
462: UserWarning: The 'normed' kwarg is deprecated, and has been replac
ed by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\chandu1\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6
462: UserWarning: The 'normed' kwarg is deprecated, and has been replac
ed by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



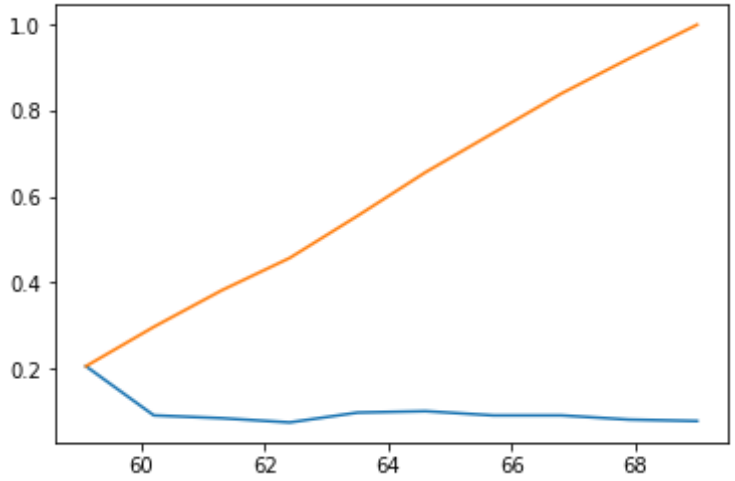
In [12]:
```
'''CDF'''

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
hm=pd.read_csv("haberman.csv")

counts, bin_edges = np.histogram(hm['age'], bins=10, density=True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:], pdf)
plt.plot(bin_edges[1:], cdf)
plt.show();
```

```
[0.05228758 0.08823529 0.1503268  0.17320261 0.17973856 0.13398693
 0.13398693 0.05882353 0.02287582 0.00653595]
[30.  35.3 40.6 45.9 51.2 56.5 61.8 67.1 72.4 77.7 83. ]
```



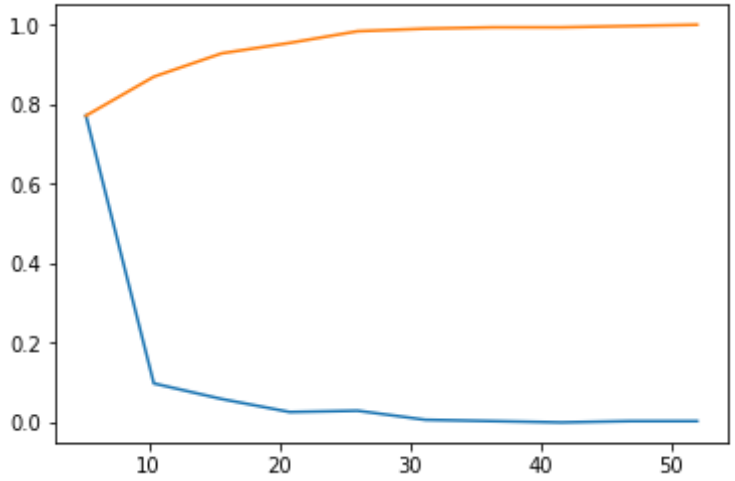
In [13]:
```
counts, bin_edges = np.histogram(hm['year'], bins=10, density=True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges);
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:], pdf)
plt.plot(bin_edges[1:], cdf)
plt.show();
```

```
[0.20588235 0.09150327 0.08496732 0.0751634  0.09803922 0.10130719
 0.09150327 0.09150327 0.08169935 0.07843137]
[58.  59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
```

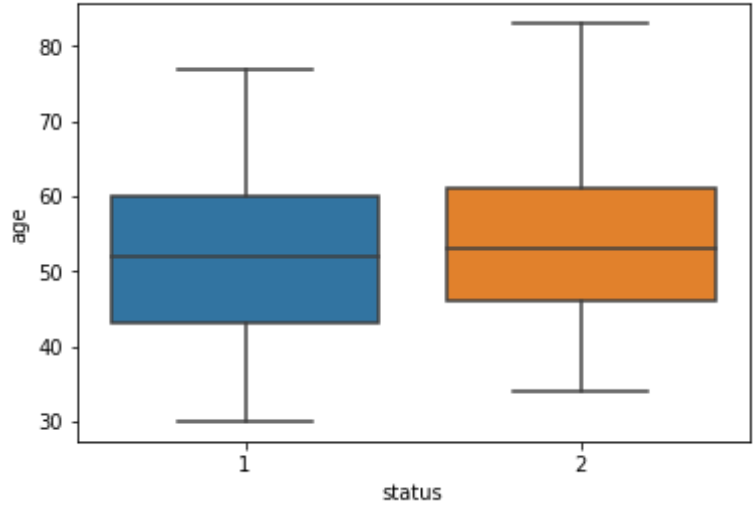


In [14]:
```
counts, bin_edges = np.histogram(hm['nodes'], bins=10, density=True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges);
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:], pdf)
plt.plot(bin_edges[1:], cdf)
plt.show();
```

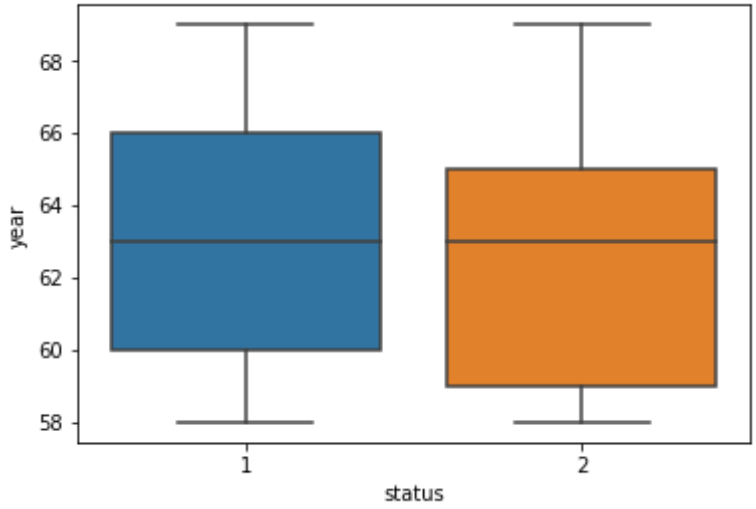
```
[0.77124183 0.09803922 0.05882353 0.02614379 0.02941176 0.00653595
 0.00326797 0.         0.00326797 0.00326797]
[ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
```



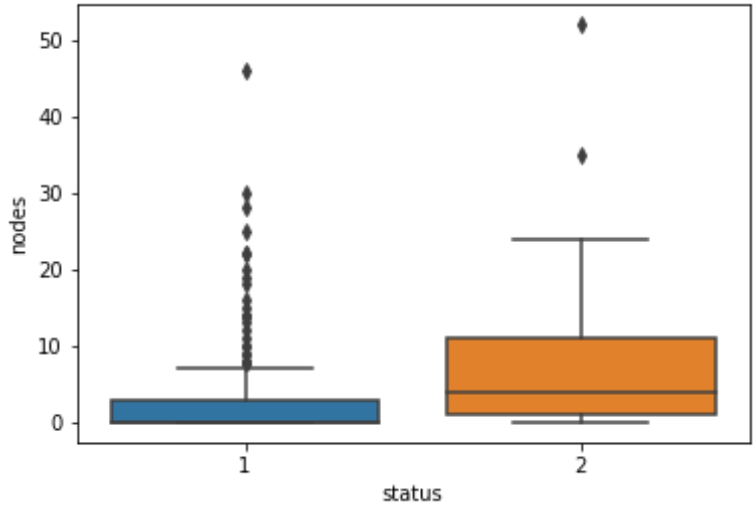
In [4]:
```
'''box plot'''
sns.boxplot(x= "status" , y="age", data=hm)
plt.show()
```



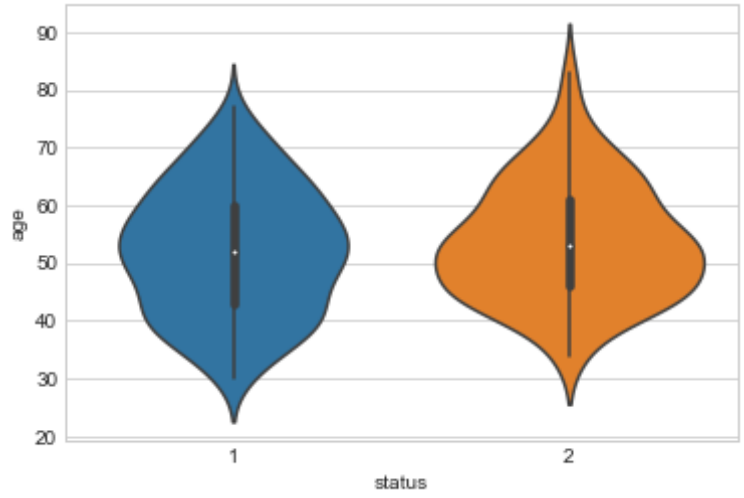
In [5]:
```
sns.boxplot(x="status" , y="year", data=hm)
plt.show()
```



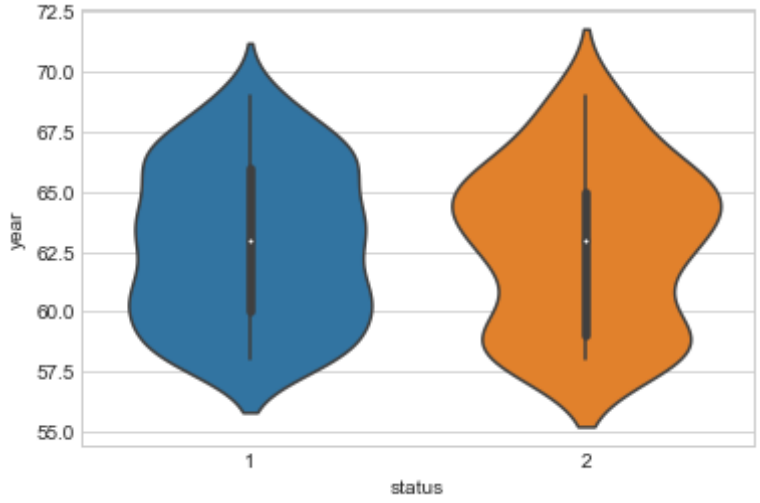
In [6]:
```
sns.boxplot(x="status" , y="nodes" , data=hm)
plt.show()
```



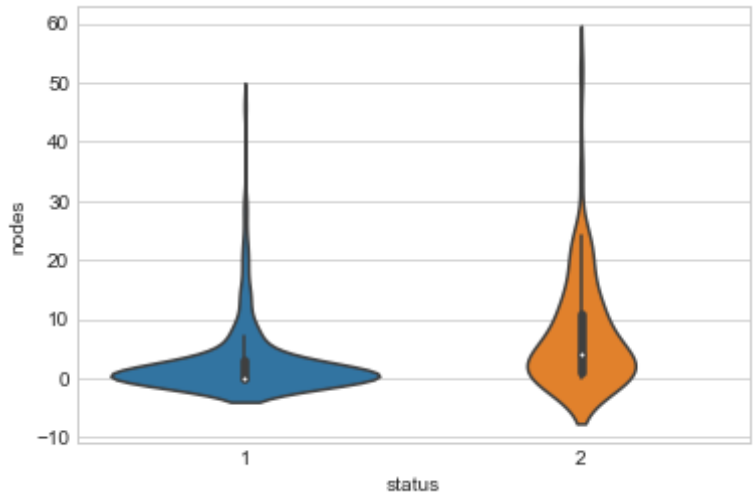
In [15]:
```
'''violin plots'''
sns.violinplot(x="status", y="age", data=hm, size=6)
plt.show()
```



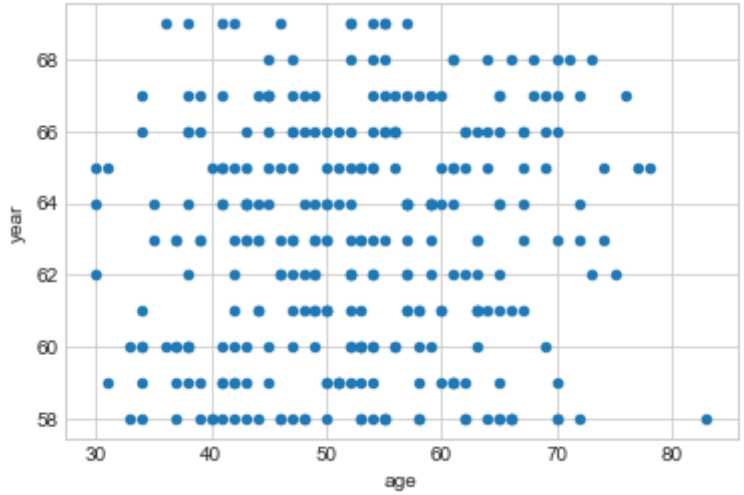
In [16]:
```
sns.violinplot(x="status", y="year", data=hm, size=6)
plt.show()
```



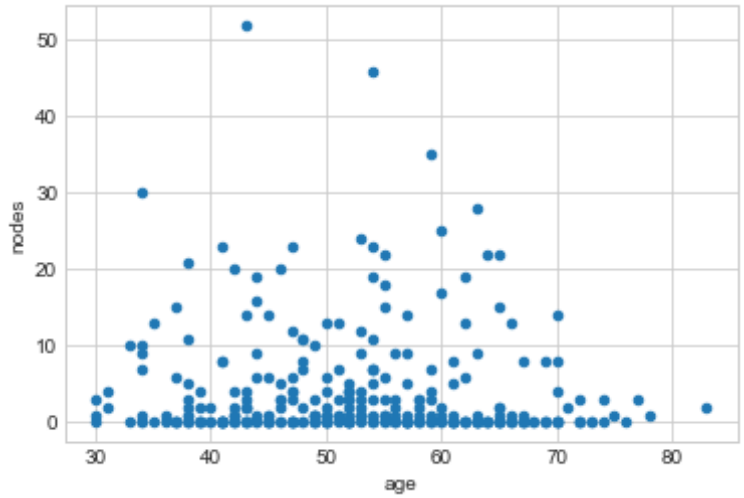
In [17]:
```
sns.violinplot(x="status", y="nodes", data=hm, size=6)
plt.show()
```



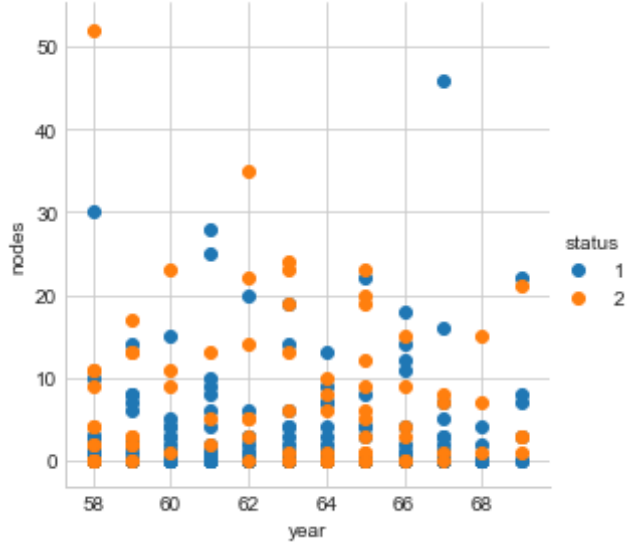
In [18]:
```
'''scatter plot'''

hm.plot(kind='scatter', x='age', y='year');
plt.show()
```



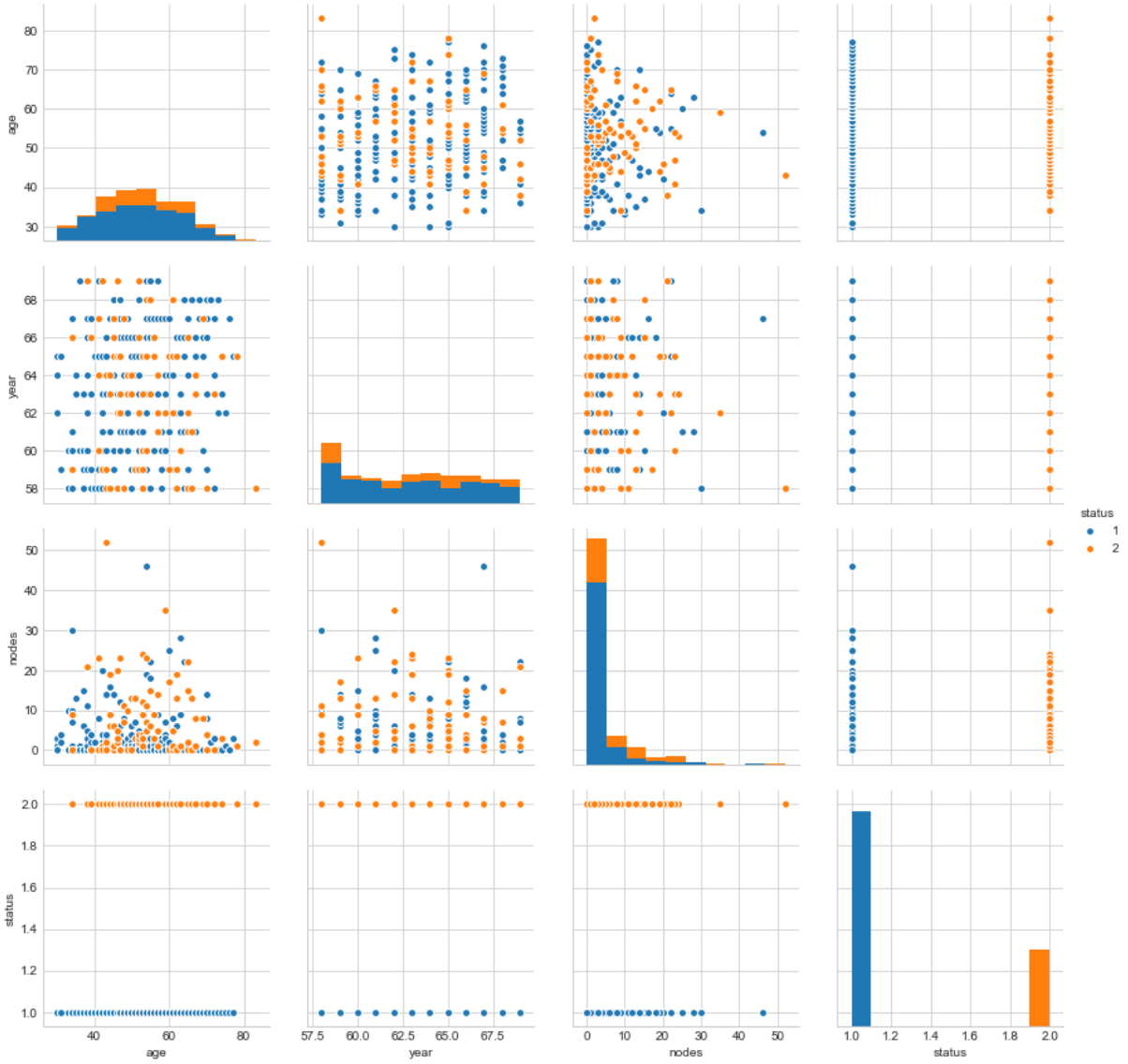
In [19]:
```
hm.plot(kind='scatter', x='age', y='nodes');
plt.show()
```



In [23]:
```
sns.set_style("whitegrid");
sns.FacetGrid(hm, hue="status", size=4) \
   .map(plt.scatter, "year", "nodes") \
   .add_legend();
plt.show();
```



In [27]:
```
'''pairplot'''

plt.close();
sns.set_style("whitegrid");
sns.pairplot(hm, hue="status", size=3);
plt.show()
```



In [ ]:
```
'''the data set clearly says that the people survived for longer than 5
 years are given the survival status as 1
which is given in blue color and the people who did not survive for mor
e than 5 years are given the survival status of
```

```
2 which is in orange color

so, by the data set we can understand that the patients vary from 30 to

83 with the median value of 52 and the maximum number of chances to sur
vive is nearly 75%
as the patients are having less nodes and 25% of the patients are havin
g more
so, this is an imbalanced data set.

the plot between year and nodes can be a better one for speration
of the two classes than the other plots.'''
```