Harberman data set

In [44]:
```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import csv

hd = pd .read_csv(r"C:\Users\raksh\Downloads\haberman data set EDA.csv")
print(hd)
```

```
     age  year  nodes  status
0     30    64      1       1
1     30    62      3       1
2     30    65      0       1
3     31    59      2       1
4     31    65      4       1
5     33    58     10       1
6     33    60      0       1
7     34    59      0       2
8     34    66      9       2
9     34    58     30       1
10    34    60      1       1
11    34    61     10       1
12    34    67      7       1
13    34    60      0       1
14    35    64     13       1
15    35    63      0       1
16    36    60      1       1
17    36    69      0       1
18    37    60      0       1
19    37    63      0       1
20    37    58      0       1
21    37    59      6       1
22    37    60     15       1
23    37    63      0       1
24    38    69     21       2
25    38    59      2       1
26    38    60      0       1
27    38    60      0       1
28    38    62      3       1
29    38    64      1       1
..   ...   ...    ...     ...
276   67    66      0       1
277   67    61      0       1
278   67    65      0       1
279   68    67      0       1
280   68    68      0       1
281   69    67      8       2
282   69    60      0       1
283   69    65      0       1
284   69    66      0       1
285   70    58      0       2
286   70    58      4       2
287   70    66     14       1
288   70    67      0       1
289   70    68      0       1
290   70    59      8       1
291   70    63      0       1
292   71    68      2       1
293   72    63      0       2
294   72    58      0       1
295   72    64      0       1
296   72    67      3       1
297   73    62      0       1
298   73    68      0       1
299   74    65      3       2
300   74    63      0       1
301   75    62      1       1
302   76    67      0       1
303   77    65      3       1
304   78    65      1       2
305   83    58      2       2

[306 rows x 4 columns]
```

In [45]:
```python
print(hd.shape)
print(hd.columns)
```

```
(306, 4)
Index(['age', 'year', 'nodes', 'status'], dtype='object')
```

In [46]:
```python
hd.describe()
```

Out[46]:

	age	year	nodes	status
count	306.000000	306.000000	306.000000	306.000000
mean	52.457516	62.852941	4.026144	1.264706
std	10.803452	3.249405	7.189654	0.441899
min	30.000000	58.000000	0.000000	1.000000
25%	44.000000	60.000000	0.000000	1.000000
50%	52.000000	63.000000	1.000000	1.000000
75%	60.750000	65.750000	4.000000	2.000000
max	83.000000	69.000000	52.000000	2.000000

In [47]:
```python
hd.isnull().sum()
```

Out[47]:
```
age       0
year      0
nodes     0
status    0
dtype: int64
```

In [48]:
```python
hd["status"].value_counts()
```

Out[48]:
```
1    225
2     81
Name: status, dtype: int64
```

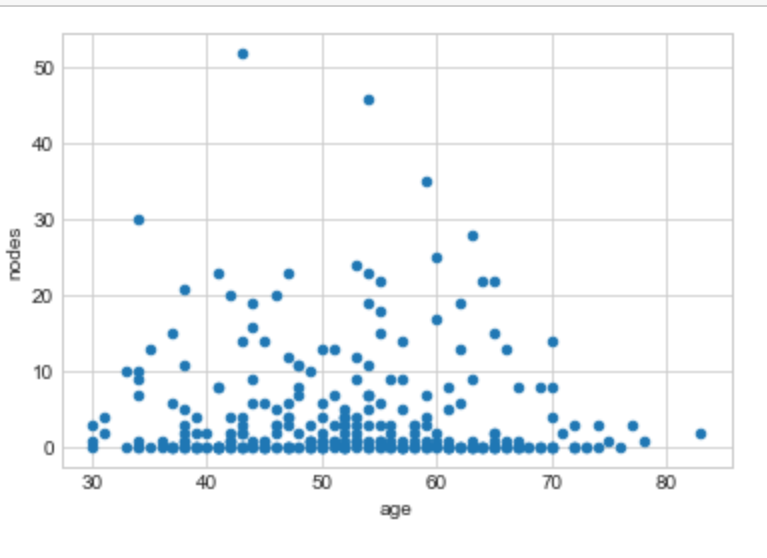
In [49]:
```python
hd['status']=hd['status'].replace(to_replace=[1,2],value=['yes','no'])
hd
```

Out[49]:

	age	year	nodes	status
0	30	64	1	yes
1	30	62	3	yes
2	30	65	0	yes
3	31	59	2	yes
4	31	65	4	yes
5	33	58	10	yes
6	33	60	0	yes
7	34	59	0	no
8	34	66	9	no
9	34	58	30	yes
10	34	60	1	yes
11	34	61	10	yes
12	34	67	7	yes
13	34	60	0	yes
14	35	64	13	yes
15	35	63	0	yes
16	36	60	1	yes
17	36	69	0	yes
18	37	60	0	yes
19	37	63	0	yes
20	37	58	0	yes
21	37	59	6	yes
22	37	60	15	yes
23	37	63	0	yes
24	38	69	21	no
25	38	59	2	yes
26	38	60	0	yes
27	38	60	0	yes
28	38	62	3	yes
29	38	64	1	yes
...	...	...	...	...
276	67	66	0	yes
277	67	61	0	yes
278	67	65	0	yes
279	68	67	0	yes
280	68	68	0	yes
281	69	67	8	no
282	69	60	0	yes
283	69	65	0	yes
284	69	66	0	yes
285	70	58	0	no
286	70	58	4	no
287	70	66	14	yes
288	70	67	0	yes
289	70	68	0	yes
290	70	59	8	yes
291	70	63	0	yes
292	71	68	2	yes
293	72	63	0	no
294	72	58	0	yes
295	72	64	0	yes
296	72	67	3	yes
297	73	62	0	yes
298	73	68	0	yes
299	74	65	3	no
300	74	63	0	yes
301	75	62	1	yes
302	76	67	0	yes
303	77	65	3	yes
304	78	65	1	no
305	83	58	2	no

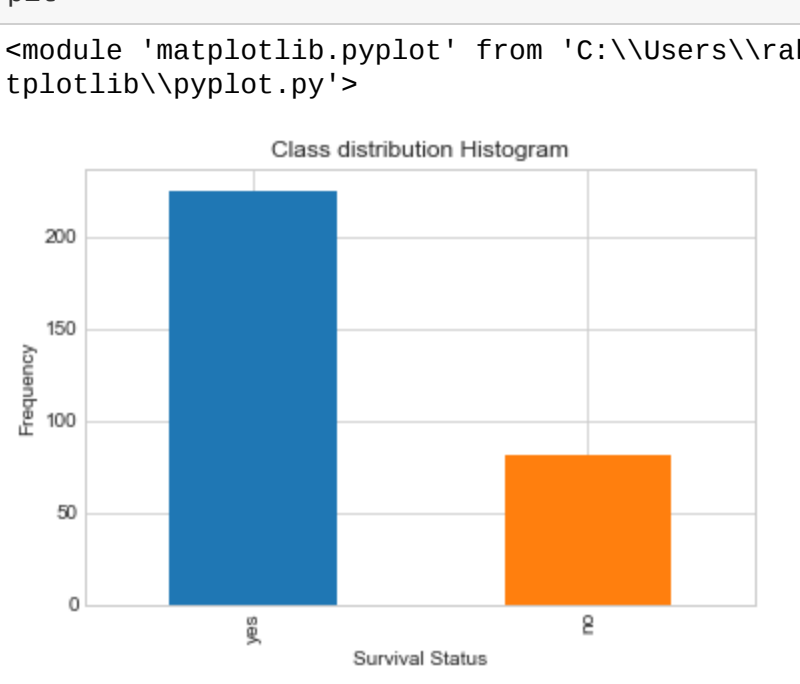
306 rows × 4 columns

In [58]:
```python
hd.plot(kind='scatter',x='age',y='nodes');
plt.show()
```



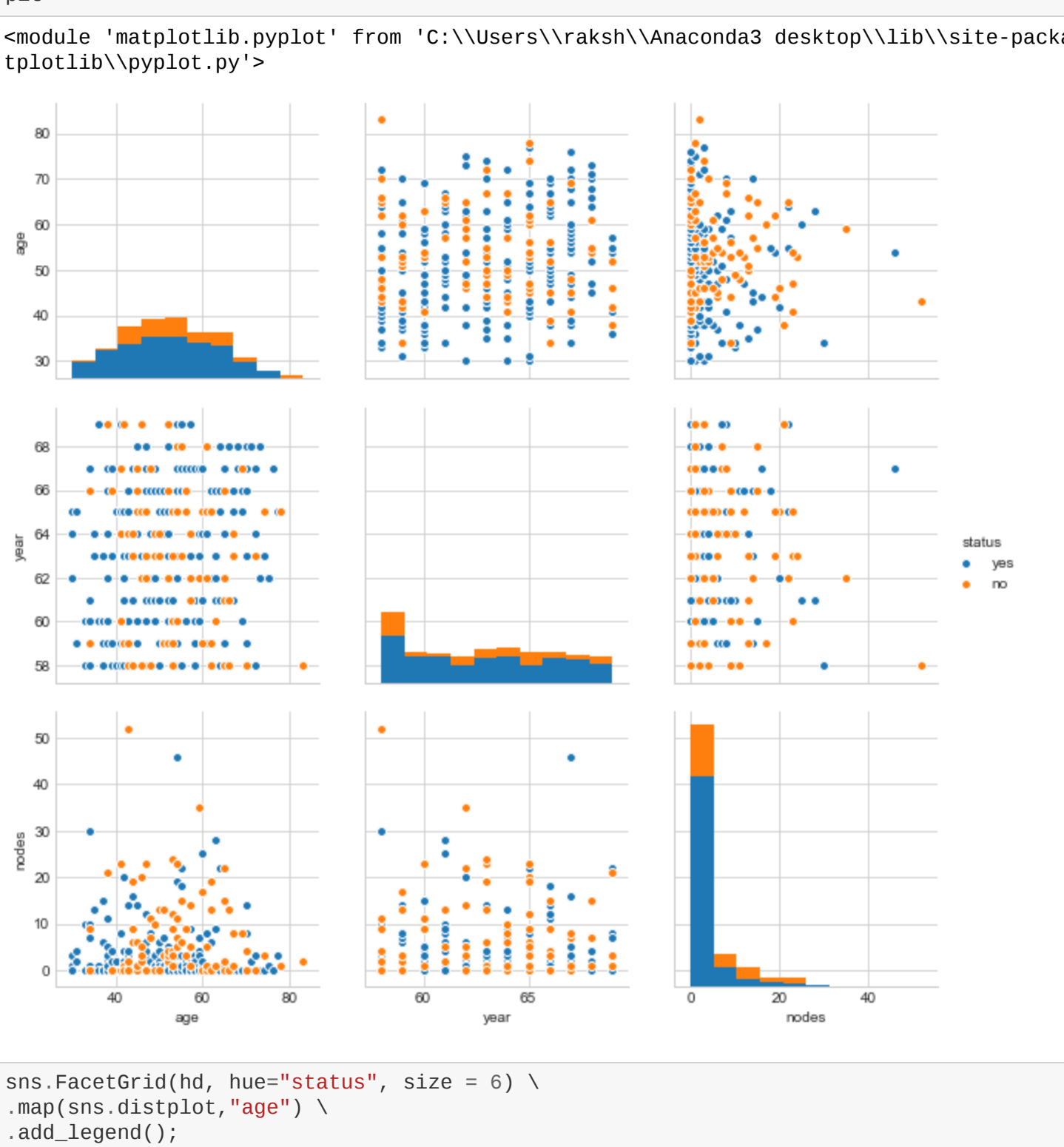
In [57]:
```python
count_status = pd.value_counts(hd["status"])
count_status.plot(kind = 'bar')
plt.title("Class distribution Histogram")
plt.xlabel("Survival Status")
plt.ylabel("Frequency")
plt
```

Out[57]: `<module 'matplotlib.pyplot' from 'C:\\Users\\raksh\\Anaconda3 desktop\\lib\\site-packages\\matplotlib\\pyplot.py'>`



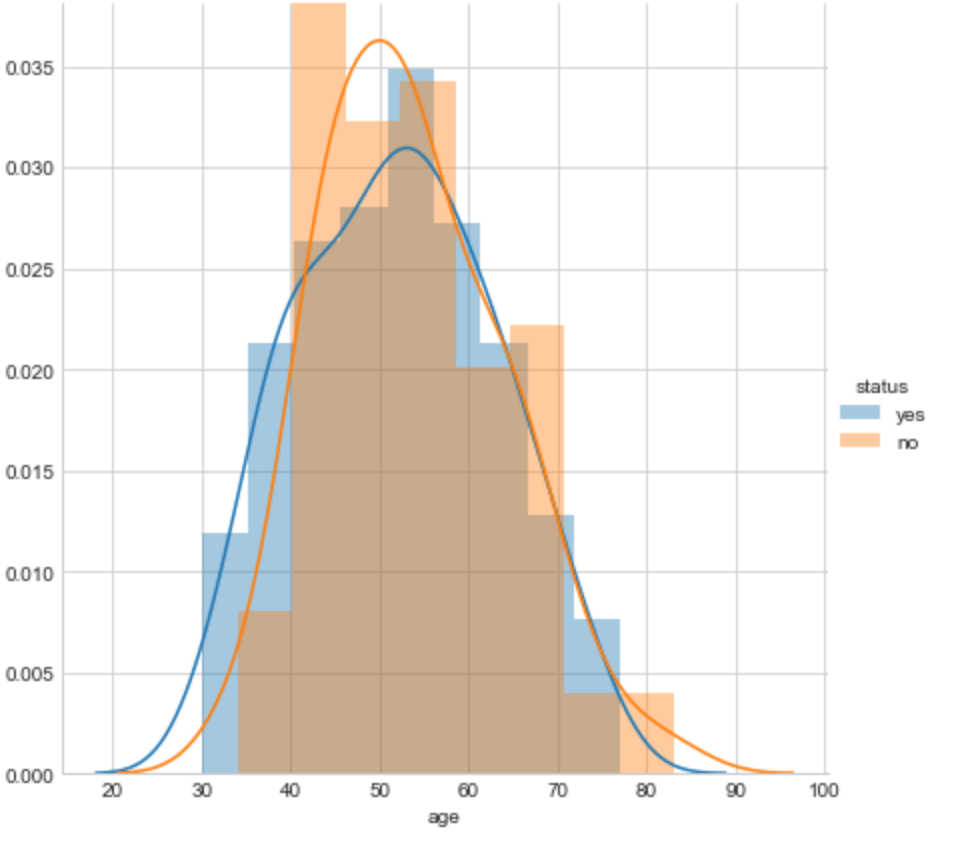
In [56]:
```python
sns.set_style("whitegrid");
sns.pairplot(hd, hue="status", size=3);
plt
```

Out[56]: `<module 'matplotlib.pyplot' from 'C:\\Users\\raksh\\Anaconda3 desktop\\lib\\site-packages\\matplotlib\\pyplot.py'>`



In [60]:
```python
sns.FacetGrid(hd, hue="status", size = 6) \
.map(sns.distplot,"age") \
.add_legend();
plt.show()
```

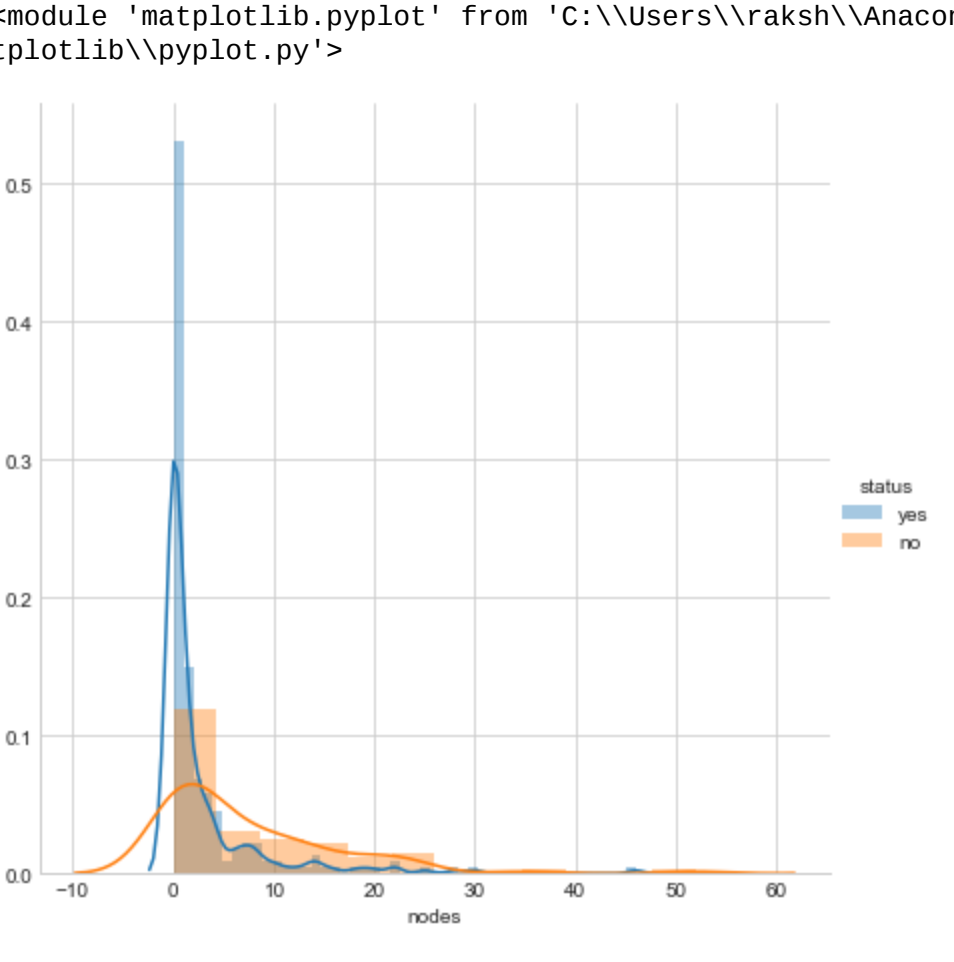
```
C:\Users\raksh\Anaconda3 desktop\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\raksh\Anaconda3 desktop\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



In [61]:
```python
sns.FacetGrid(hd, hue="status", size = 6) \
.map(sns.distplot,"nodes") \
.add_legend();
plt
```

```
C:\Users\raksh\Anaconda3 desktop\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\raksh\Anaconda3 desktop\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

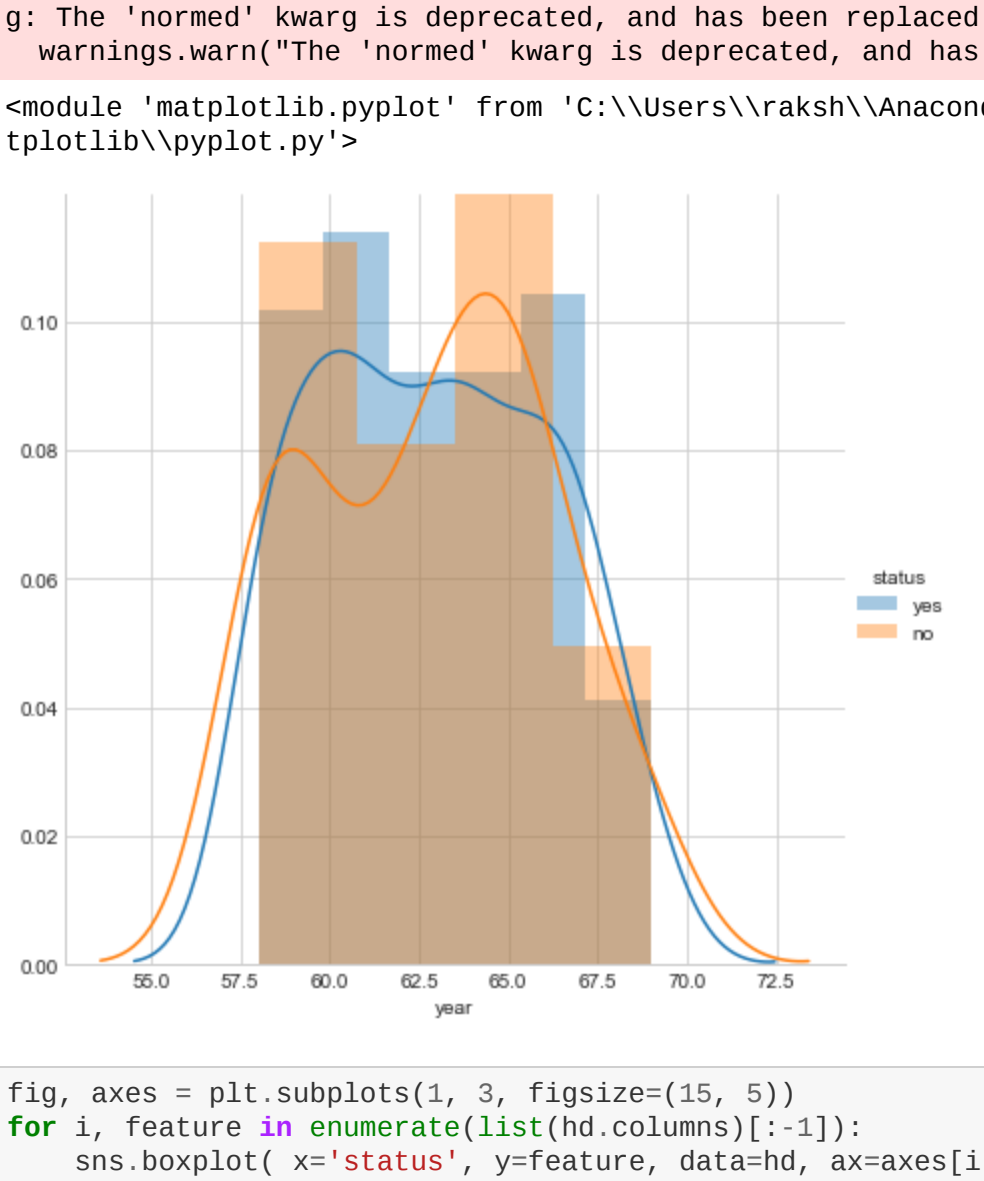
Out[61]: `<module 'matplotlib.pyplot' from 'C:\\Users\\raksh\\Anaconda3 desktop\\lib\\site-packages\\matplotlib\\pyplot.py'>`



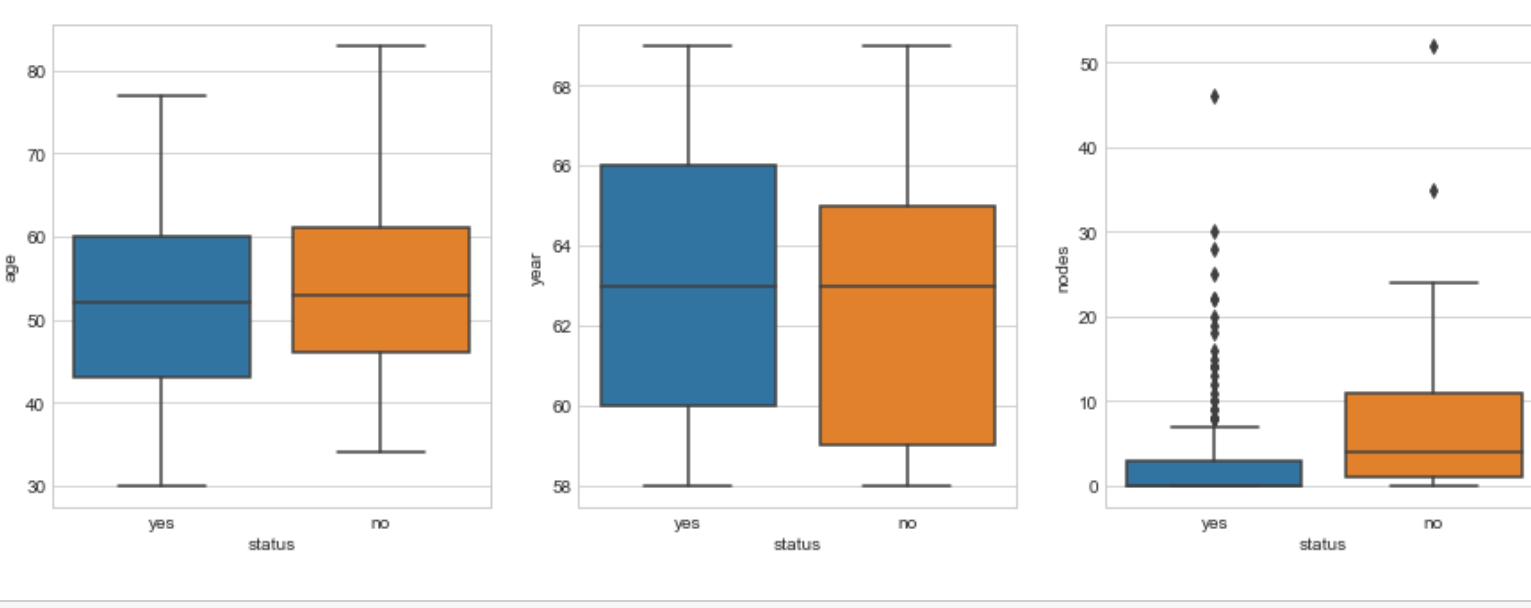
In [64]:
```python
sns.FacetGrid(hd, hue="status", size = 6) \
.map(sns.distplot,"year") \
.add_legend();

plt
```

```
C:\Users\raksh\Anaconda3 desktop\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\raksh\Anaconda3 desktop\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

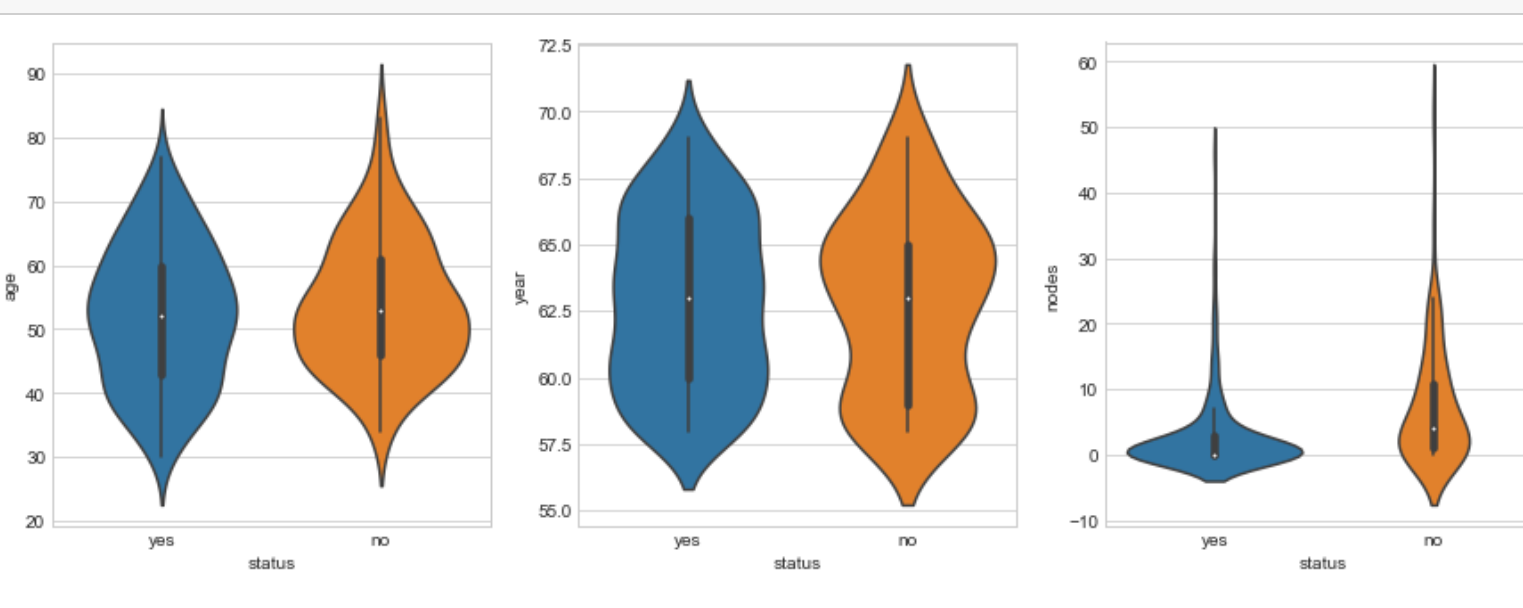
Out[64]: `<module 'matplotlib.pyplot' from 'C:\\Users\\raksh\\Anaconda3 desktop\\lib\\site-packages\\matplotlib\\pyplot.py'>`



In [63]:
```python
fig, axes = plt.subplots(1, 3, figsize=(15, 5))
for i, feature in enumerate(list(hd.columns)[:-1]):
    sns.boxplot( x='status', y=feature, data=hd, ax=axes[i])
plt.show()
```



In [65]:
```python
fig, axes = plt.subplots(1, 3, figsize=(15, 5))
for i, feature in enumerate(list(hd.columns)[:-1]):
    sns.violinplot( x='status', y=feature, data=hd, ax=axes[i])
plt.show()
```



In [68]:
```python
plt.figure(figsize=(20,5))
for i, feature in enumerate(list(hd.columns)[:-1]):
    plt.subplot(1, 3, i+1)
    print(feature+" *********")
    counts, bin_edges = np.histogram(hd[feature], bins=10, density=True)
    print("Bin Edges: {}".format(bin_edges))
    pdf = counts/sum(counts)
    print("PDF: {}".format(pdf))
    cdf = np.cumsum(pdf)
    print("CDF: {}".format(cdf))
    plt.plot(bin_edges[1:], pdf, bin_edges[1:], cdf)
    plt.xlabel(feature)
```

```
age *********
Bin Edges: [30.   35.3 40.6 45.9 51.2 56.5 61.8 67.1 72.4 77.7 83. ]
PDF: [0.05228758 0.08823529 0.1503268  0.17320261 0.17973856 0.13398693
 0.13398693 0.05882353 0.02287582 0.00653595]
CDF: [0.05228758 0.14052288 0.29084967 0.46405229 0.64379085 0.77777778
 0.91176471 0.97058824 0.99346405 1.        ]
year *********
Bin Edges: [58.   59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
PDF: [0.20588235 0.09150327 0.08496732 0.0751634  0.09803922 0.10130719
 0.09150327 0.09150327 0.08169935 0.07843137]
CDF: [0.20588235 0.29738562 0.38235294 0.45751634 0.55555556 0.65686275
 0.74836601 0.83986928 0.92156863 1.        ]
nodes *********
Bin Edges: [ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
PDF: [0.77124183 0.09803922 0.05882353 0.02614379 0.02941176 0.00653595
 0.00326797 0.         0.00326797 0.00326797]
CDF: [0.77124183 0.86928105 0.92810458 0.95424837 0.98366013 0.99019608
 0.99346405 0.99346405 0.99673203 1.        ]
```