In [1]:

```
'''This document is for analysis of cancer patients with the intention of getting some
 data insights'''
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
#Loading the csv file
haberman= pd.read_csv("C:\\Users\\ashwani\\Desktop\\tab\\Applied course\\haberman\\habe
rman.csv")
```

In [3]:

```
haberman.head()
```

Out[3]:

	Age	Year	Axil Nodes	Survival 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

We have four features, survival status is the label. Year of operarion does not play any important role in deciding the result so we've got two features i.e. age and number of axil nodes.

In [4]:

```
#maximum number of nodesfound in a patient
haberman["Axil Nodes"].max()
```

Out[4]:

```
52
```

In [5]:

```
#survival status for max number of axil nodes
haberman[haberman["Axil Nodes"]==haberman["Axil Nodes"].max()][["Age","Survival Status"
]]
```

Out[5]:

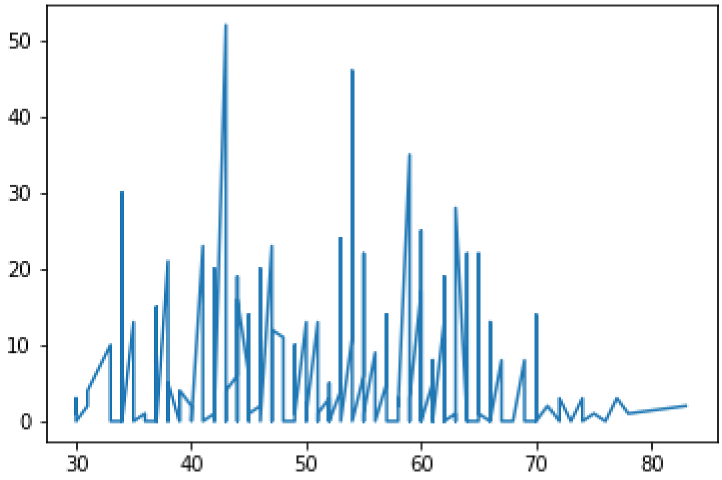
	Age	Survival Status
62	43	2

In [6]:

```
fig=plt.figure(1)
#plt.subplot(1,2,1)
plt.plot(haberman["Age"],haberman["Axil Nodes"])
```

Out[6]:

```
[<matplotlib.lines.Line2D at 0xe0e0f50>]
```



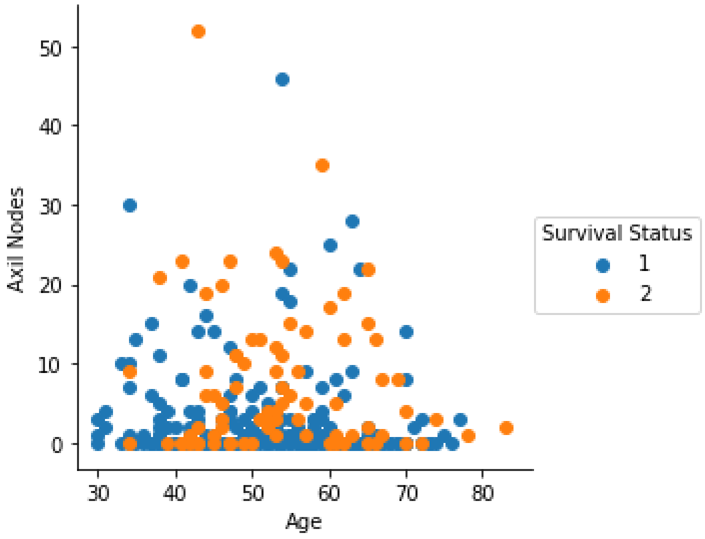
Number of axil nodes are lower iin the patients over the ag eof 70years.

In [7]:

```
sns.FacetGrid(data=haberman,hue="Survival Status",size=4)\
     .map(plt.scatter,"Age","Axil Nodes").add_legend()
```

Out[7]:

```
<seaborn.axisgrid.FacetGrid at 0xd07b8f0>
```



In [8]:

```
haberman["Axil Nodes"].value_counts()
```

Out[8]:

```
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: Axil Nodes, dtype: int64
```

In [9]:

```
#distribution in the classes
haberman["Survival Status"].value_counts()
```

Out[9]:

```
1    225
2     81
Name: Survival Status, dtype: int64
```

No. of survivors are more than no. of non-survivors for more than five years.

In [10]:

```
haberman.shape
```

Out[10]:

```
(306, 4)
```

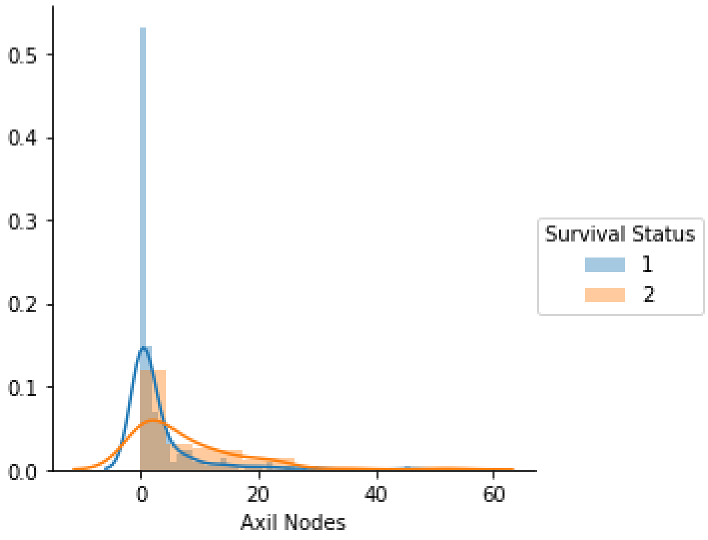
In [11]:

```
sns.FacetGrid(data= haberman,hue="Survival Status" ,size=4)\
    .map(sns.distplot, "Axil Nodes")\
    .add_legend()
```

```
c:\users\ashwani\appdata\local\programs\python\python36-32\lib\site-packag
es\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is depre
cated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

Out[11]:

```
<seaborn.axisgrid.FacetGrid at 0xe161b10>
```



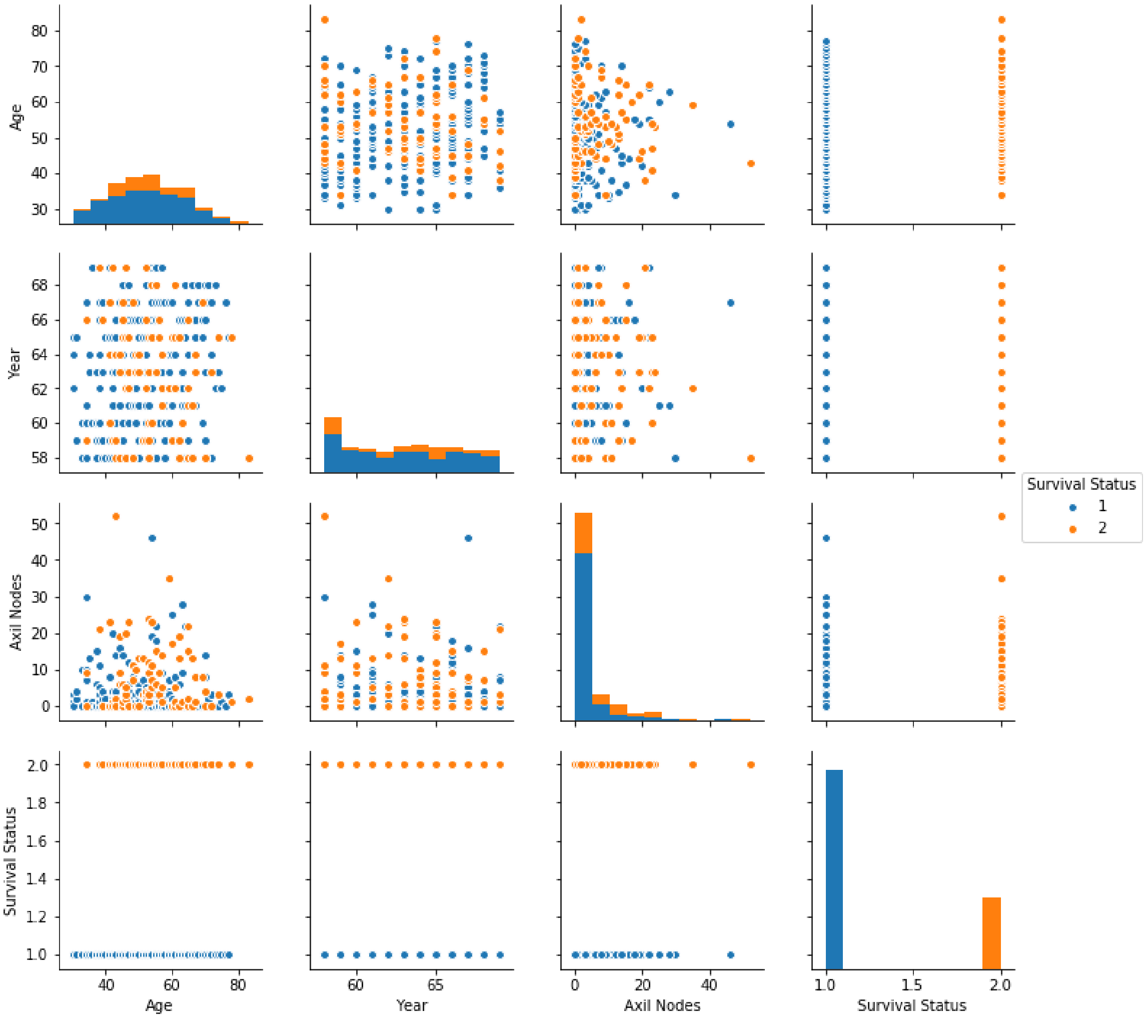
most people have 0 axil nodes. variance for survivors is more than that of non-survivors

In [12]:

```
sns.pairplot(haberman,hue="Survival Status")
```

Out[12]:

```
<seaborn.axisgrid.PairGrid at 0xe12c410>
```



In [13]:

```
#saved=haberman.loc[haberman["Survival Status"]==1]
#notsaved=haberman.loc[haberman["Survival Status"]==2]
```

In [14]:

```
counts,binedges= np.histogram(haberman["Axil Nodes"],bins=10,density= True)
```

In [15]:

```
pdf=counts/sum(counts)
pdf
```

Out[15]:

```
array([0.77124183, 0.09803922, 0.05882353, 0.02614379, 0.02941176,
       0.00653595, 0.00326797, 0.        , 0.00326797, 0.00326797])
```

In [16]:

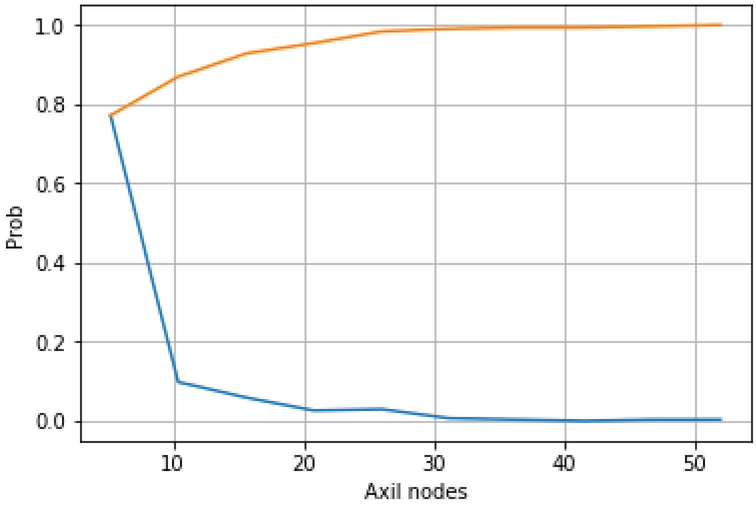
```
cdf=np.cumsum(pdf)
cdf
```

Out[16]:

```
array([0.77124183, 0.86928105, 0.92810458, 0.95424837, 0.98366013,
       0.99019608, 0.99346405, 0.99346405, 0.99673203, 1.        ])
```

In [17]:

```
plt.plot(binedges[1:],pdf)
plt.plot(binedges[1:],cdf)
plt.xlabel("Axil nodes")
plt.ylabel("Prob")
plt.grid()
```



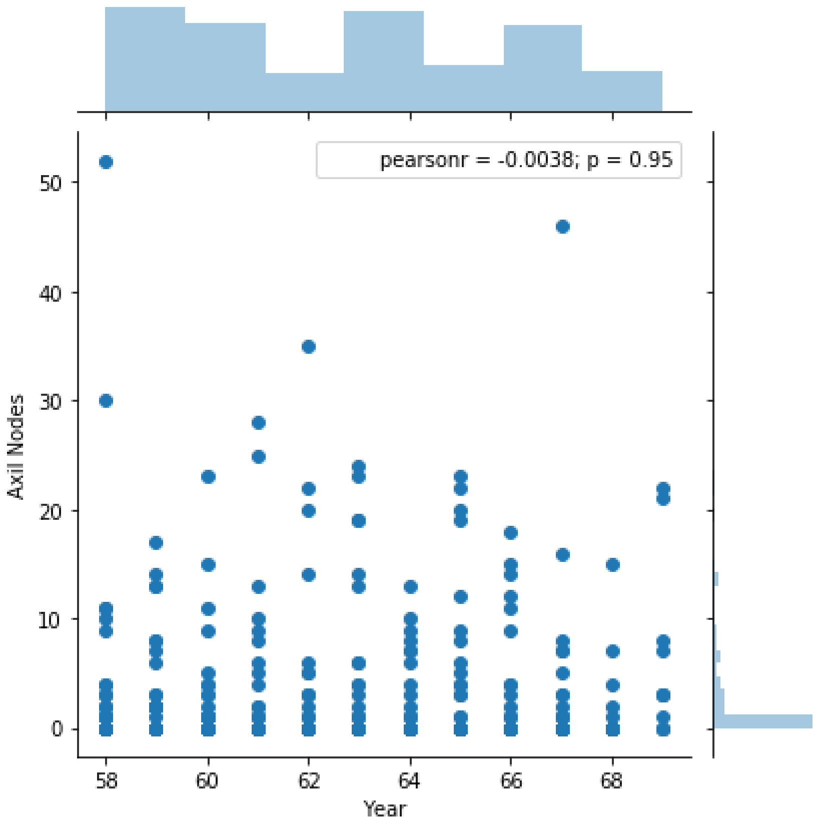
In [23]:

```
sns.jointplot(haberman["Year"],haberman["Axil Nodes"])
```

```
c:\users\ashwani\appdata\local\programs\python\python36-32\lib\site-packag
es\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed' kwarg is depre
cated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

Out[23]:

```
<seaborn.axisgrid.JointGrid at 0x1255f50>
```

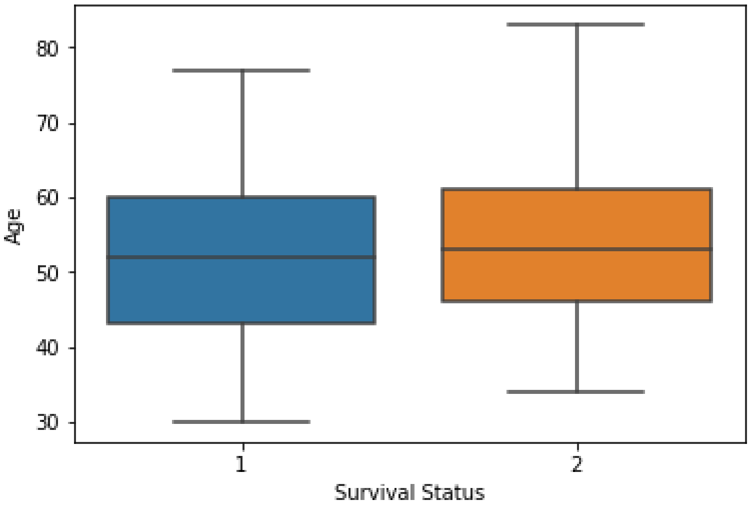


In [19]:

```
sns.boxplot(data=haberman,x="Survival Status",y="Age")
```

Out[19]:

```
<matplotlib.axes._subplots.AxesSubplot at 0xf4f9b90>
```

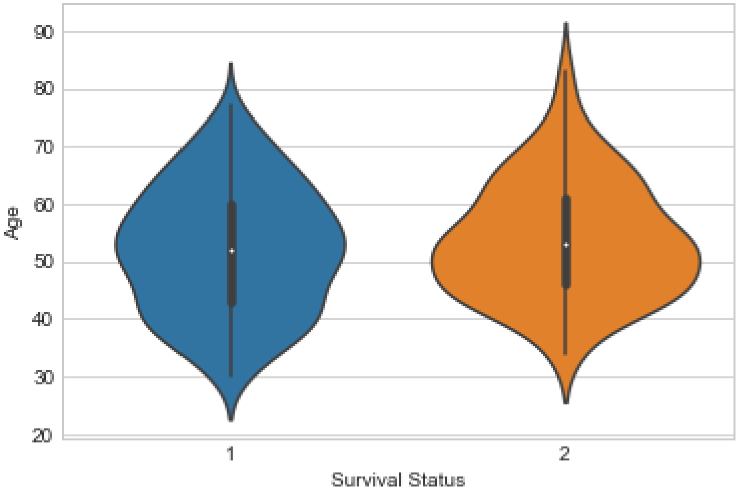


In [24]:

```
sns.set_style("whitegrid")
sns.violinplot(data=haberman,x="Survival Status",y="Age")
```

Out[24]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1626bd0>
```



chances of survival are slightly greater after the age of 50.