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

In [3]:
```
Haberman = pd.read_csv('haberman.csv')
```

In [5]:
```
print(Haberman.columns)
```

```
Index(['30', '64', '1', '1.1'], dtype='object')
```

In [6]:
```
Haberman.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 305 entries, 0 to 304
Data columns (total 4 columns):
30     305 non-null int64
64     305 non-null int64
1      305 non-null int64
1.1    305 non-null int64
dtypes: int64(4)
memory usage: 9.6 KB
```

In [7]:
```
Haberman.columns=['Age','pa_year_oper','pos_aux_nodes','Patients_survived_more
5years']
```

In [8]:
```
Haberman['Patients_survived_more5years']=Haberman['Patients_survived_more5year
s'].map({1:"Yes",2:"No"})
```

In [9]:
```
Haberman.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 305 entries, 0 to 304
Data columns (total 4 columns):
Age                            305 non-null int64
pa_year_oper                   305 non-null int64
pos_aux_nodes                  305 non-null int64
Patients_survived_more5years    305 non-null object
dtypes: int64(3), object(1)
memory usage: 9.6+ KB
```

Observation

Dataset is contaning 4 columns with Age,patients year of operation, auxiliary nodes and how many years they survived. Each column in the data set is having 305Rows without null value.

In [12]:

```
Haberman.head()
```

Out[12]:

	Age	pa_year_oper	pos_aux_nodes	Patients_survived_more5years
0	30	62	3	Yes
1	30	65	0	Yes
2	31	59	2	Yes
3	31	65	4	Yes
4	33	58	10	Yes

In [13]:

```
Haberman['Patients_survived_more5years'].value_counts()
```

Out[13]:

```
Yes    224
No      81
Name: Patients_survived_more5years, dtype: int64
```

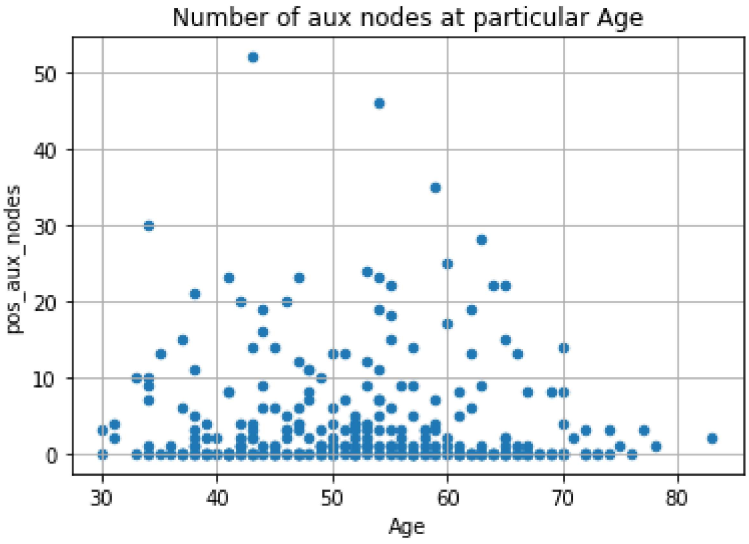
Observation

224 patients have survived more than 5 years after the operation.

81 patients have survived less than 5 years after the operation.

In [16]:

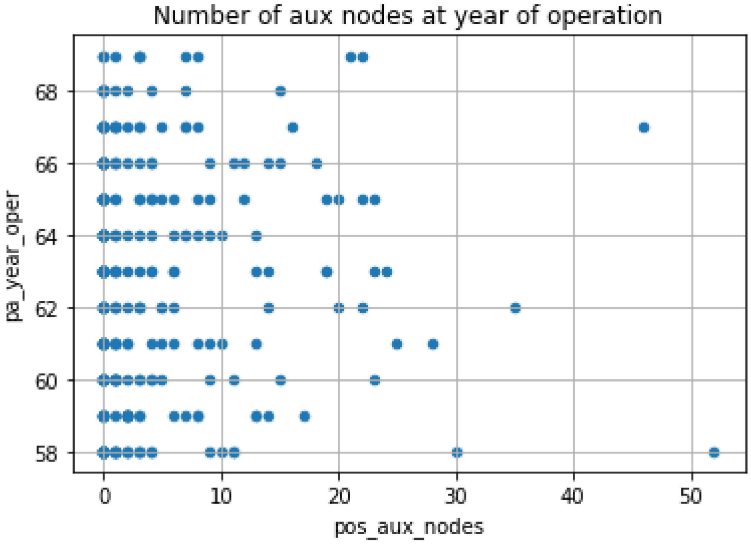
```
Haberman.plot(kind = 'scatter', x='Age',y='pos_aux_nodes')
plt.grid()
plt.title('Number of aux nodes at particular Age')
plt.show()
```



Observation

Most of the auxiliary nodes are in between 0 to 10 for all age groups.

In [19]:
```
Haberman.plot(kind = 'scatter', x='pos_aux_nodes',y='pa_year_oper')
plt.grid()
plt.title('Number of aux nodes at year of operation')
plt.show()
```



Observation

Maximum number of nodes are in between 0 to 10 at the year of operation.

In [20]:
```
Haberman[:][Haberman['pos_aux_nodes']==Haberman['pos_aux_nodes'].max()]
```

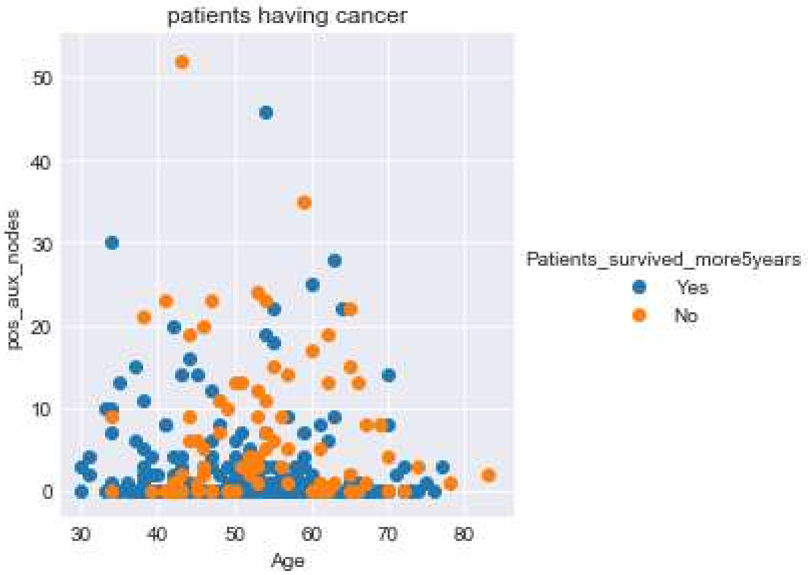
Out[20]:

	Age	pa_year_oper	pos_aux_nodes	Patients_survived_more5years
61	43	58	52	No

Observation

maximum number of pos_aux_nodes are encountered in the year 1958

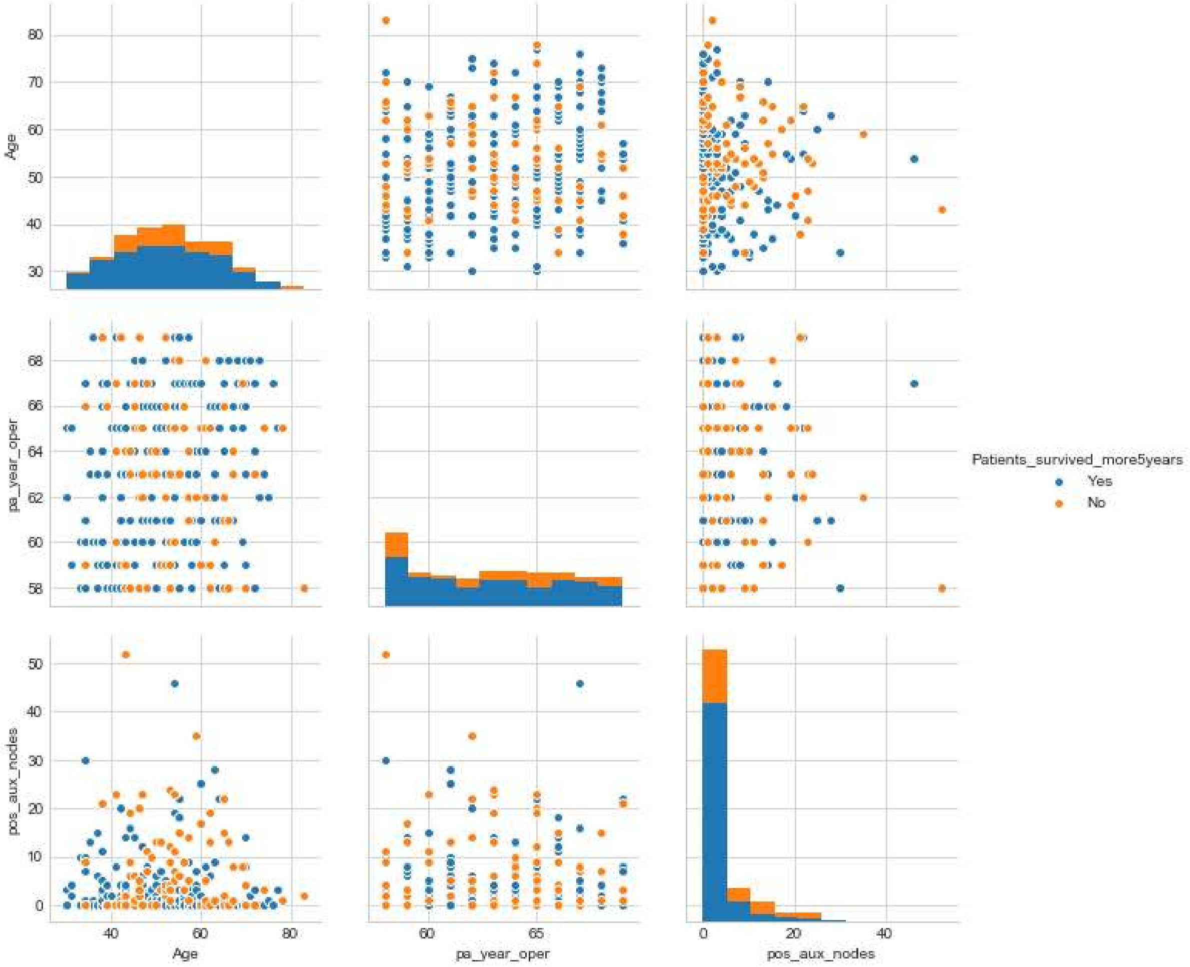
In [26]:
```
sns.set_style('darkgrid')
sns.FacetGrid(Haberman,hue = 'Patients_survived_more5years',size=4).map(plt.sc
atter,'Age','pos_aux_nodes').add_legend()
plt.title('patients having cancer')
plt.show()
```



Observation

Most of the patients who survived more than five years are having positive auxiliary nodes below 10.

```
In [28]: plt.close()
         sns.set_style('whitegrid')
         sns.pairplot(Haberman,hue = 'Patients_survived_more5years',size=3)
         plt.show()
```



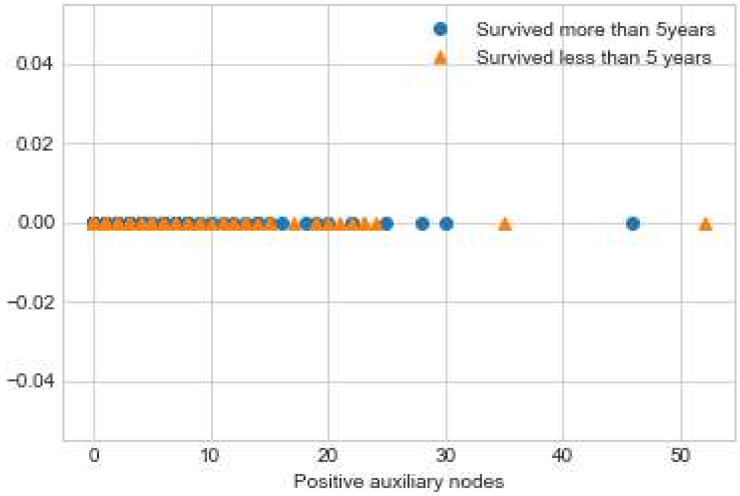
observation

From the above plots we are not able to separate the patients who has survived more than five years and who had not survived. As most of the patients are having positive auxiliary nodes below 10.

```
In [30]: Haberman_yes=Haberman.loc[Haberman['Patients_survived_more5years']=='Yes']
```

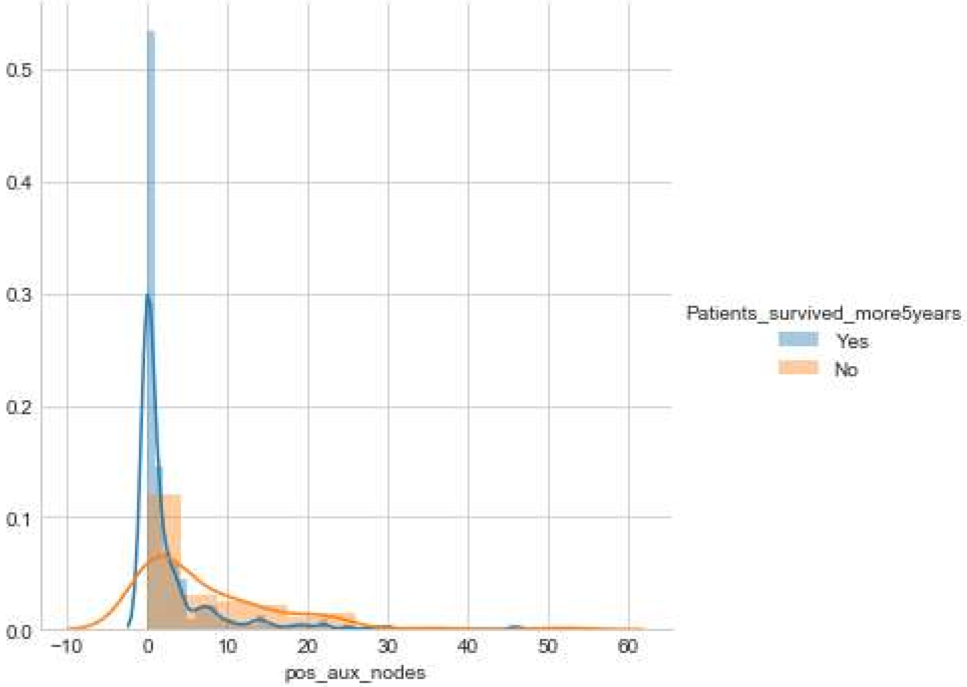
```
In [31]: Haberman_no=Haberman.loc[Haberman['Patients_survived_more5years']=='No']
```

In [35]:
```
plt.plot(Haberman_yes['pos_aux_nodes'],np.zeros_like(Haberman_yes['pos_aux_nodes']),'o',label='Survived more than 5years')
plt.plot(Haberman_no['pos_aux_nodes'],np.zeros_like(Haberman_no['pos_aux_nodes']),'^',label='Survived less than 5 years')
plt.legend()
plt.xlabel("Positive auxiliary nodes")
plt.show()
```



In [36]:
```
import warnings
warnings.filterwarnings("ignore")
sns.FacetGrid(Haberman,hue='Patients_survived_more5years',size=5).map(sns.distplot,'pos_aux_nodes').add_legend()
```

Out[36]:
```
<seaborn.axisgrid.FacetGrid at 0x1ef45bdb160>
```

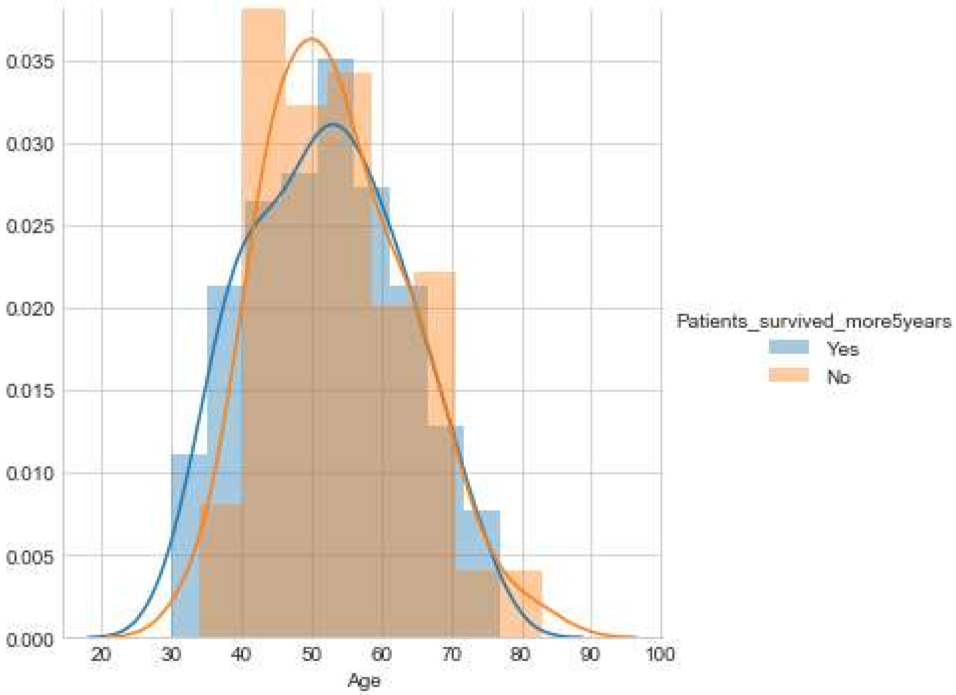


observation

we can say that most of the patients who survived more than five years having positive auxillary nodes between 0 to 7 .

```
In [37]: sns.FacetGrid(Haberman,hue='Patients_survived_more5years',size=5).map(sns.dist
         plot,'Age').add_legend()
```

```
Out[37]: <seaborn.axisgrid.FacetGrid at 0x1ef44d6a208>
```

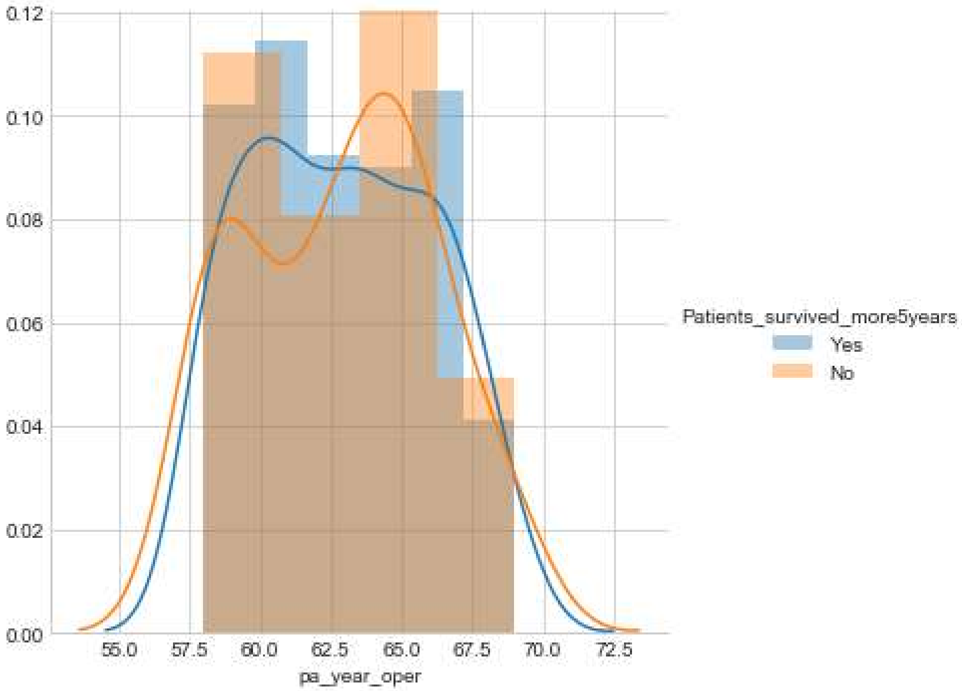


In [38]:

```
sns.FacetGrid(Haberman,hue='Patients_survived_more5years',size=5).map(sns.dist
plot,'pa_year_oper').add_legend()
```

Out[38]:

```
<seaborn.axisgrid.FacetGrid at 0x1ef44eb7668>
```



In [39]:

```
Haberman.describe()
```

Out[39]:

	Age	pa_year_oper	pos_aux_nodes
count	305.000000	305.000000	305.000000
mean	52.531148	62.849180	4.036066
std	10.744024	3.254078	7.199370
min	30.000000	58.000000	0.000000
25%	44.000000	60.000000	0.000000
50%	52.000000	63.000000	1.000000
75%	61.000000	66.000000	4.000000
max	83.000000	69.000000	52.000000

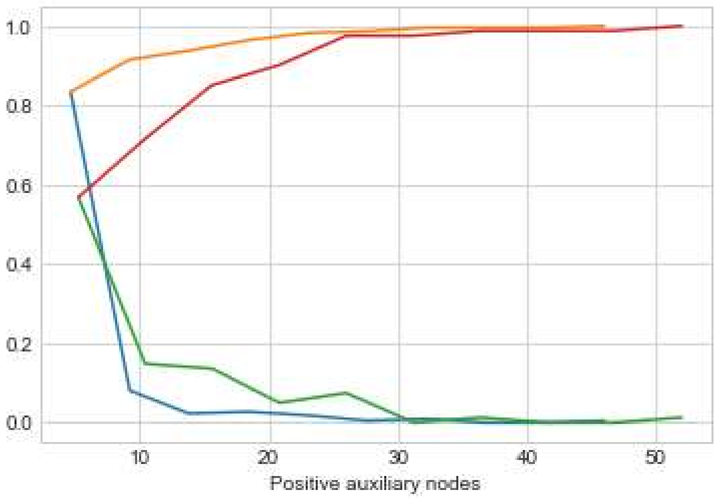
In [40]:

```
counts, bin_edges = np.histogram(Haberman_yes['pos_aux_nodes'], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)

counts,bin_edges = np.histogram(Haberman_no['pos_aux_nodes'],bins=10,density=T
rue)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.xlabel('Positive auxiliary nodes')
```

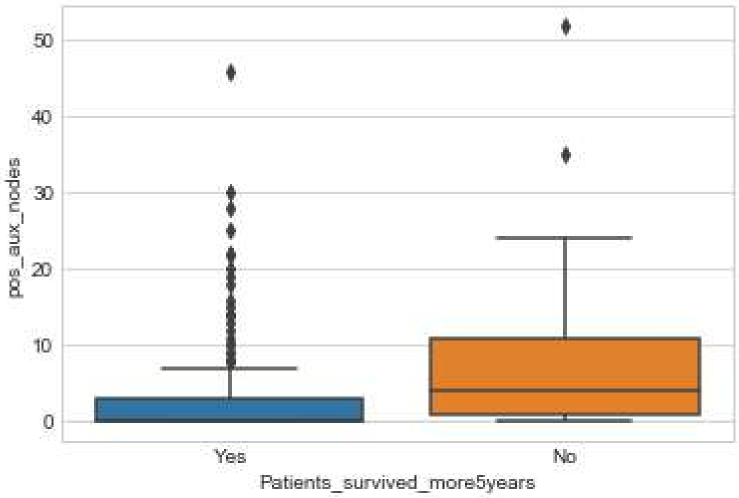
Out[40]:

```
Text(0.5,0,'Positive auxiliary nodes')
```



In [41]:

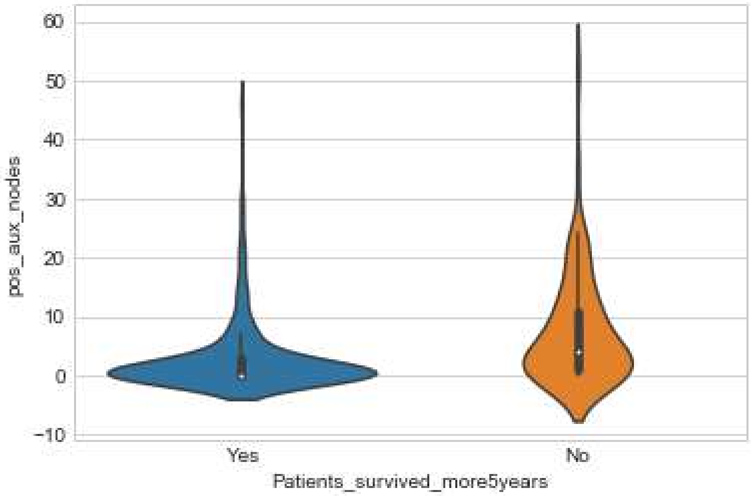
```
sns.boxplot(y='pos_aux_nodes',x='Patients_survived_more5years',data=Haberman)
plt.show()
```



observation

from the above boxplot we can say that 75% people survived more than five years having auxiliary nodes less than 4.

```
In [139]: sns.violinplot(x='Patients_survived_more5years', y='pos_aux_nodes',data=Haberm
          an)
          plt.show()
```



```
In [144]: print(np.percentile(Haberman_yes['pos_aux_nodes'],np.arange(0,100,25)))
          print(np.percentile(Haberman_no['pos_aux_nodes'],np.arange(0,100,25)))

          [0. 0. 0. 3.]
          [ 0.  1.  4. 11.]
```

```
In [147]: print(np.percentile(Haberman_yes['pos_aux_nodes'],90))
          print(np.percentile(Haberman_no['pos_aux_nodes'],90))

          8.0
          20.0
```

Objective

How many patients having cancer survived more than five years.

Conculsion

224 patients have survived more than five years after operation.

81 patients have survived less than five years after operation.

90% of the patients who survived more than five years have the positive auxiliary nodes 8.

90% of the patients who died before five years have positive auxiliary nodes 20.