In [1]:

```
import os
os.getcwd()
```

Out[1]:

```
'C:\\Users\\prem\\ML_COLLECTION\\HABERMAN_ASSIGNMENT'
```

In [2]:

```
print(os.listdir('C:/Users/prem/ML_COLLECTION/HABERMAN_ASSIGNMENT'))
```

```
['.ipynb_checkpoints', 'haberman.csv', 'haberman_assignment.ipynb', 'haberman_data
set.zip']
```

In [3]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
sns.set()
```

In [4]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()

table = pd.read_csv('C:/Users/prem/ML_COLLECTION/HABERMAN_ASSIGNMENT/haberman.csv', he
print(table.head(10))
```

```
   AgeOf_patient  year_of_operation  no_of_positive_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
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
```

As shown above , the first column describes age of patients second column describes year of operation third colummn describes positive auxillary nodes (int type data) fourth column describes survival status of patient (the survival chances of patients after 5 years) ---> if survival status value is 1 ,patient survived for more than 5 years ---> if survival status value is 2 ,patient survived for less than 5 years

In [5]:

```
print(table.describe())   # High level statistics of the dataset
```

```
       AgeOf_patient  year_of_operation  no_of_positive_nodes  survival_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
```

OBSERVATIONS:

COUNT: The total no of patients present in the dataset

-----> The total no of patients observed is 306

MEAN: The mean values of each column

-----> The mean value of ages observed is 52 years and survival status is 1.2(which means there are more no. of patients survived after 5 years)

STD: It gives the standard deviation of each field.Tells about the dispersion that is observed in the each data point

MIN : shows the minimum value in the each column

-----> The minimum age of patient is found out to be 30 years and minimum year of operation is observed to be 58

25,50,75th Percentile : shows this particular percentages of all the values in particular columnare less than this value

-----> 25% of all the patients age is observed to be less than 44 years

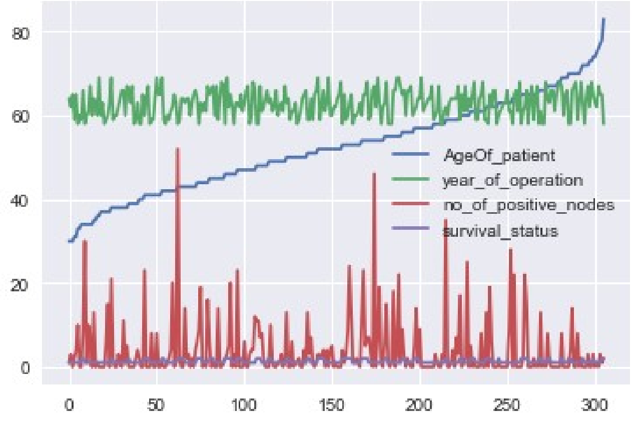
-----> 50% of all the patients age is observed to be less than 52 years

-----> 75% of all the patients age is observed to be less than 60 years

MAX : The maximum value in a column

-----> maximum age of any patient is observed to be 83 years

In [6]:
```
table.plot()          #
plt.show()
```



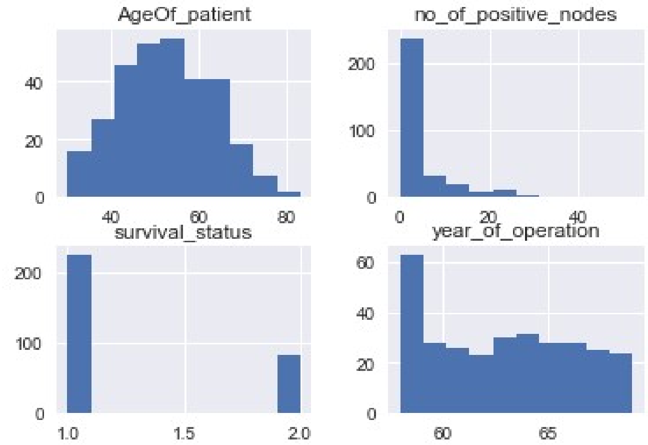
Observations :

The year of operation streches around 60 with zigzag pattern

Random spikes are observed in the plot related to positive nodes

In [7]:

```
table.hist()
plt.show()
```



OBSERVATIONS:

Age of patient plot : It is observed that more number of patients are from age around 50-60

no of positive nodes : more no of patients(>200) are with less no of positive nodes

survival status : survival status is more for 200+ patients

year of operation : bar height is maximum at early years.small variations are observed after 1960 which denotes constant number of patients got treated.

In [8]:

```
table['survival_status'].value_counts()
```

Out[8]:

```
1    225
2     81
Name: survival_status, dtype: int64
```

as shown above , the survival status which means the patients survived for more than 5 years was 225 out of 306

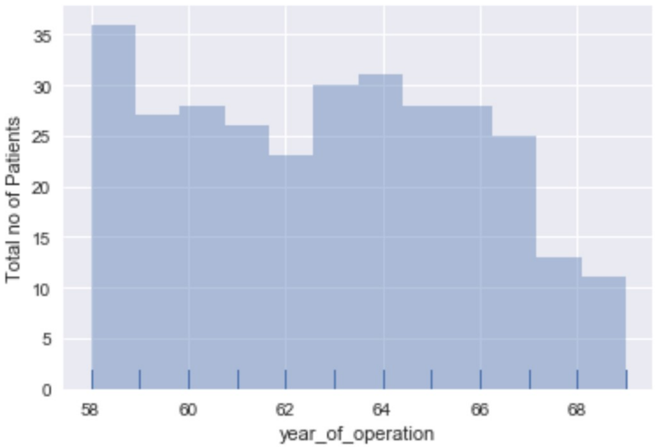
and the patients who survived for less than 5 years was 81 out of 306

In [9]:

```
# univariate analysis(PDF,CDF, boxplot,violin plot)
sns.set(color_codes=True)
%matplotlib inline
sns.distplot(table['year_of_operation'],bins=12,kde=False,rug=True).set(ylabel='Total
```

Out[9]:

```
[Text(0,0.5,'Total no of Patients')]
```



Considering the observation from above plot,

--->more no. of patients are observed at 1958

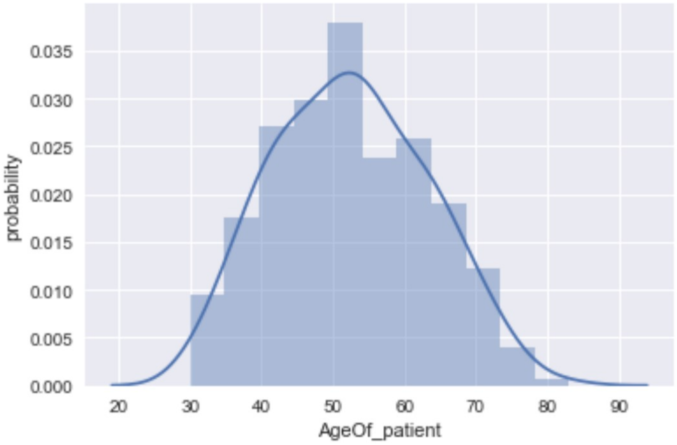
In [10]:

```

sns.distplot(table['AgeOf_patient']).set(ylabel='probability')
```

Out[10]:

```
[Text(0,0.5,'probability')]
```

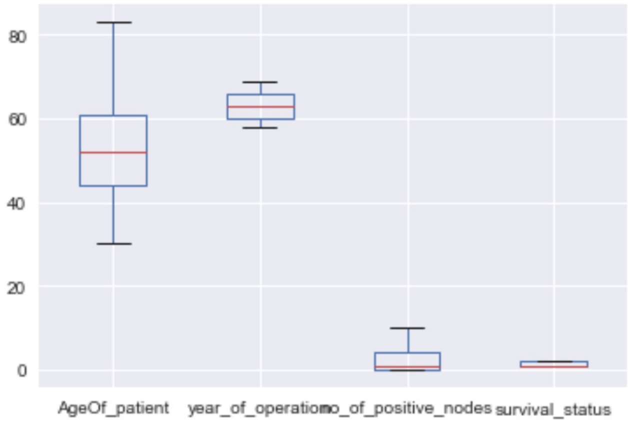


Observations:

--->As shown above, the probability of a person having age around 50 is more since the peak of distribution is at that position

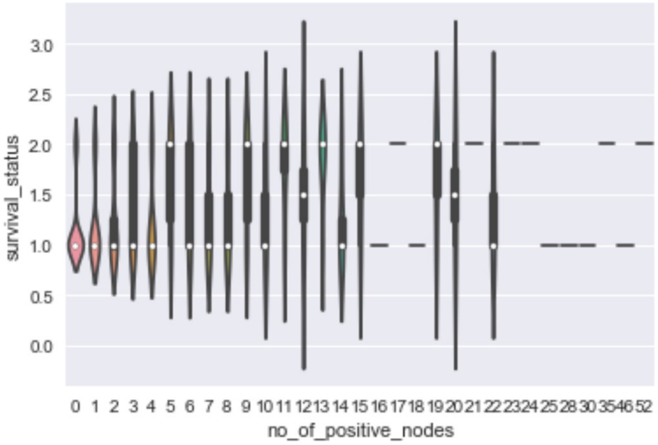
In [11]:

```
table.plot(kind='box')
plt.show()
```



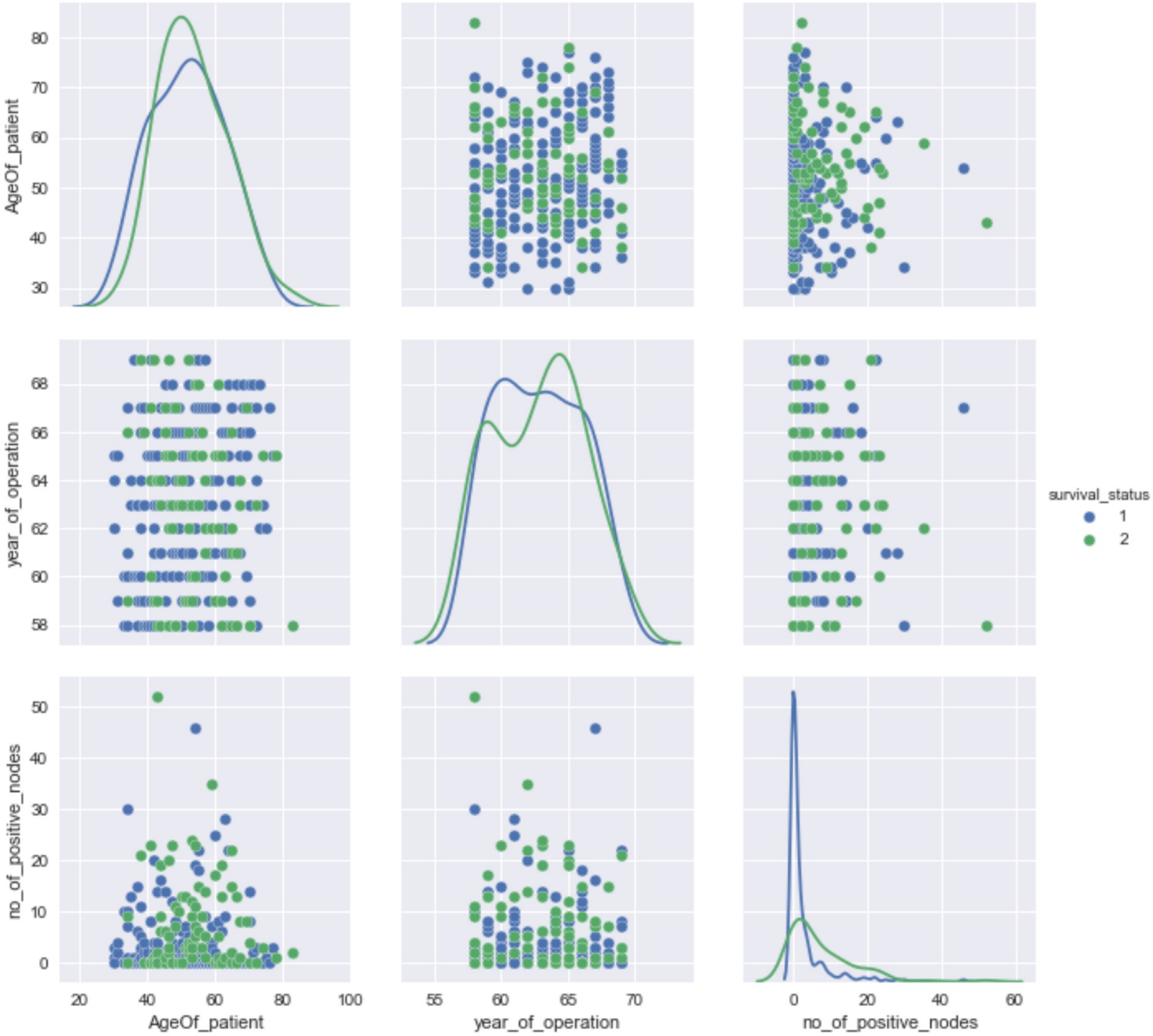
In [17]:

```
# violin plot
sns.violinplot(x="no_of_positive_nodes", y="survival_status", data=table, size=8)
plt.show()
```



OBSERVATIONS: Considering data from violin plot, if there are less no of auxillary nodes the chance of survival for more than 5 years is more (make a look at first 10 points of positive nodes , 50th percentile at most of plots remains at 1)

In [20]:
```
# pairplot
sns.pairplot(table,vars = ['AgeOf_patient', 'year_of_operation', 'no_of_positive_nodes
plt.show()
```



CONCLUSION:

----> Age of patient dosen't have effect on survival status

----> year of operation dosen't have any effect on survival status

----> Number of positive nodes has effect on survival status(less no of positive nodes , more the chances of survival after 5 years)

In [ ]: