In [260]:
```
#Haberman data set from kaggle.com
```

In [261]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
#Haberman's Survival Data Set
#Survival of patients who had undergone surgery for breast cancer
haberman=pd.read_csv("haberman.csv",names=['Patient age','year of operation','ax nodes','survivalsta
tus'])
print(haberman)
#how many data points and features
print(haberman.shape)
```

```
     Patient age  year of operation  ax nodes  survivalstatus
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]
(306, 4)
```

In [231]:
```
#print the name of columns
print(haberman.columns)
```

```
Index(['Patient age', 'year of operation', 'ax nodes', 'survivalstatus'], dtype='object')
```

In [232]:
```
#in which age the patient admitted for operaton(print the age and year of opration)
haberman[['Patient age','year of operation']]
```

Out[232]:

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

306 rows × 2 columns

In [233]:
```
print(haberman.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306 entries, 0 to 305
Data columns (total 4 columns):
Patient age          306 non-null int64
year of operation    306 non-null int64
ax nodes             306 non-null int64
survivalstatus       306 non-null int64
dtypes: int64(4)
memory usage: 9.6 KB
None
```

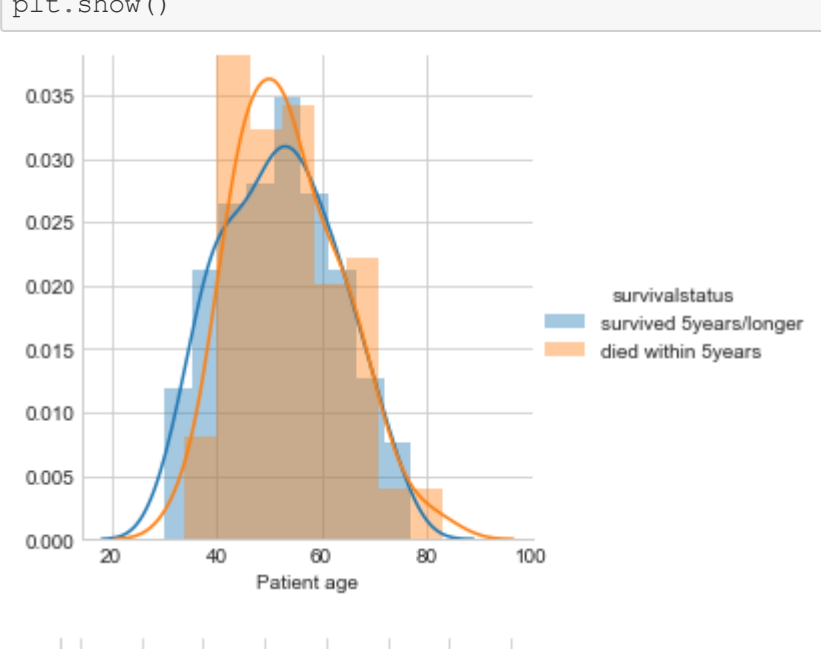
In [234]:
```
#survival status
haberman=pd.read_csv("haberman.csv",names=['Patient age','year of operation','ax nodes','survivalsta
tus'])
haberman['survivalstatus']=haberman['survivalstatus'].map({1:'survived 5years/longer',2:'died within
 5years'})
print(haberman)
#balanced dataset or imbalanced dataset according to survivalstatus
print(haberman["survivalstatus"].value_counts())
```

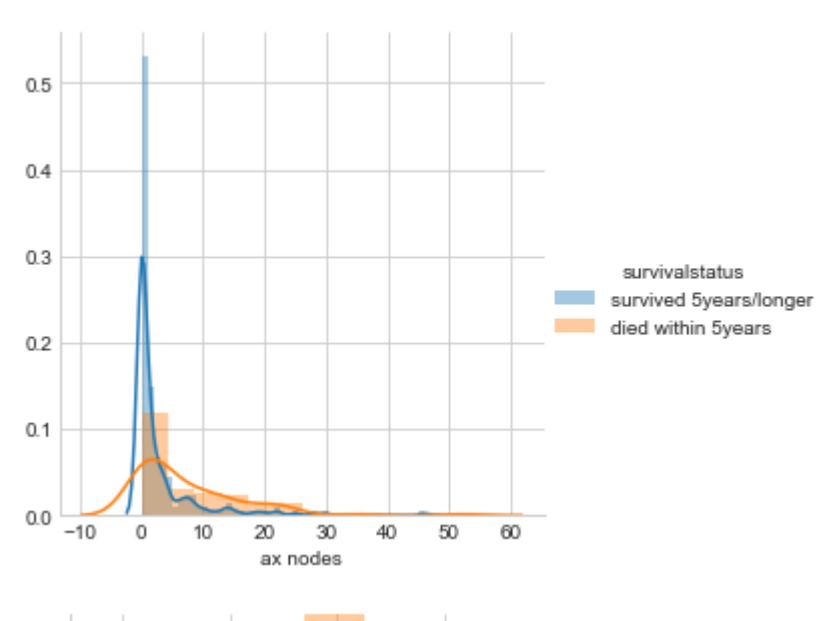
```
     Patient age  year of operation  ax nodes          survivalstatus
0             30                 64         1  survived 5years/longer
1             30                 62         3  survived 5years/longer
2             30                 65         0  survived 5years/longer
3             31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

[306 rows x 4 columns]
survived 5years/longer    225
died within 5years         81
Name: survivalstatus, dtype: int64
```

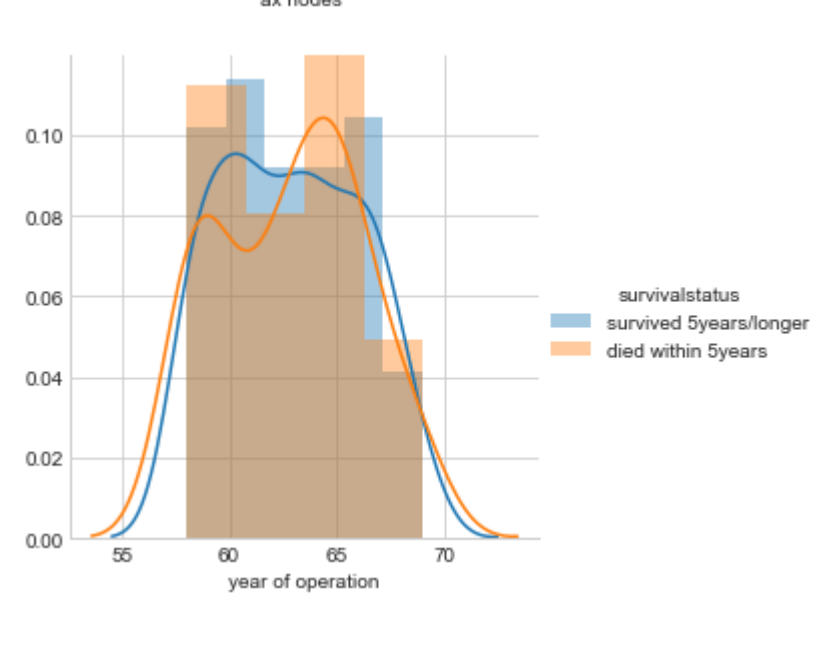
Observation- 1.In haberman dataframe 306 rows,and 4 columns is present. 2.columns names are-['Patient age','year of operation','ax nodes','survivalstatus']. 3.observed that survivalstatus has 2 different class of Patients- among all the patients 225 number of patients can survived 5 years or longer & 81 Patients died within 5 years so. 4.it is Imbalanced data set.

Objective --> Our objective is analyse the data to figure out the Patients survivalstatus based upon the Patient age,year of operation & ax nodes.

In [235]:
```
#1.Univariate Analysis
#PDF
import warnings
warnings.filterwarnings("ignore")
sns.FacetGrid(haberman,hue="survivalstatus",size=4)\
.map(sns.distplot,"Patient age")\
.add_legend()
sns.FacetGrid(haberman,hue="survivalstatus",size=4)\
.map(sns.distplot,"ax nodes")\
.add_legend()
sns.FacetGrid(haberman,hue="survivalstatus",size=4)\
.map(sns.distplot,"year of operation")\
.add_legend()
plt.show()
```







Observation-1.PDF is the smoothed version of the histogram.

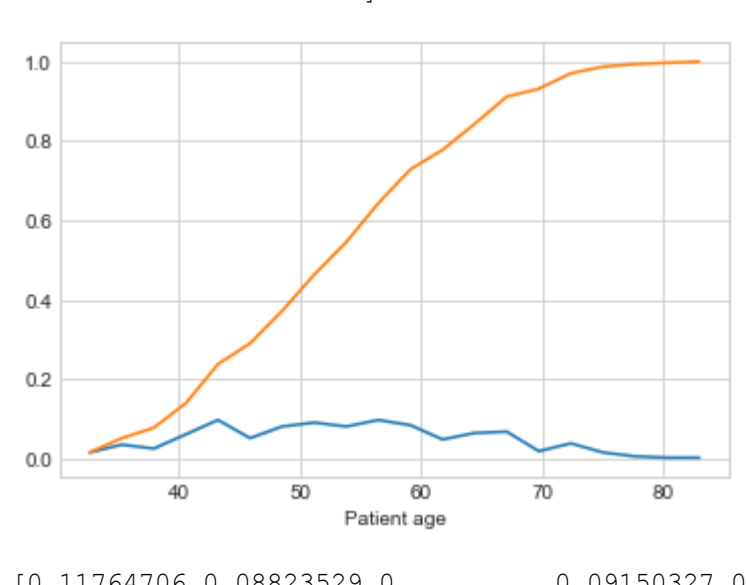
2.plots are highly overlapped,little bit difficult to identify by PDF.

3.It is imbalanced data set.

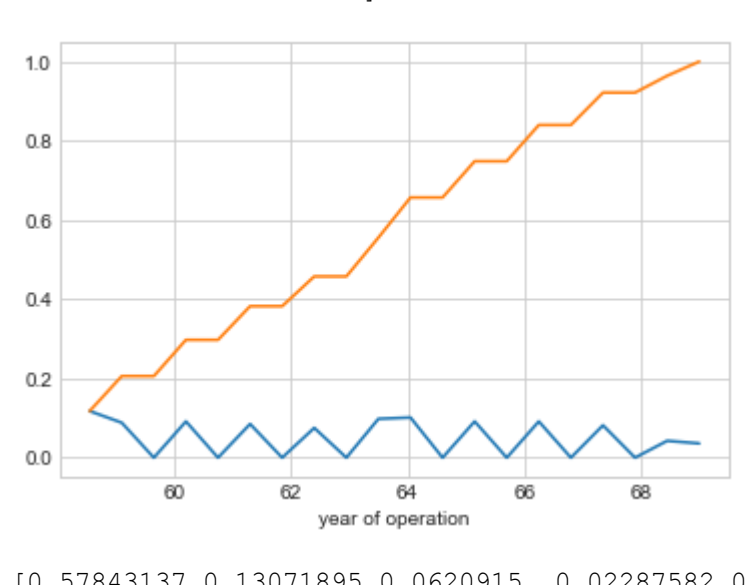
4.In ax nodes-both distribution are little different from each other.

In [248]:
```
#CDF
import numpy as np
counts,bin_edges=np.histogram(haberman['Patient age'],bins=20,density=True)
pdf=counts/sum(counts)
print(pdf)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.xlabel("Patient age")
plt.show()
counts,bin_edges=np.histogram(haberman['year of operation'],bins=20,density=True)
pdf=counts/sum(counts)
print(pdf)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.xlabel("year of operation")
plt.show()
counts,bin_edges=np.histogram(haberman['ax nodes'],bins=30,density=True)
pdf=counts/sum(counts)
print(pdf)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.xlabel("ax nodes")
plt.show()
```

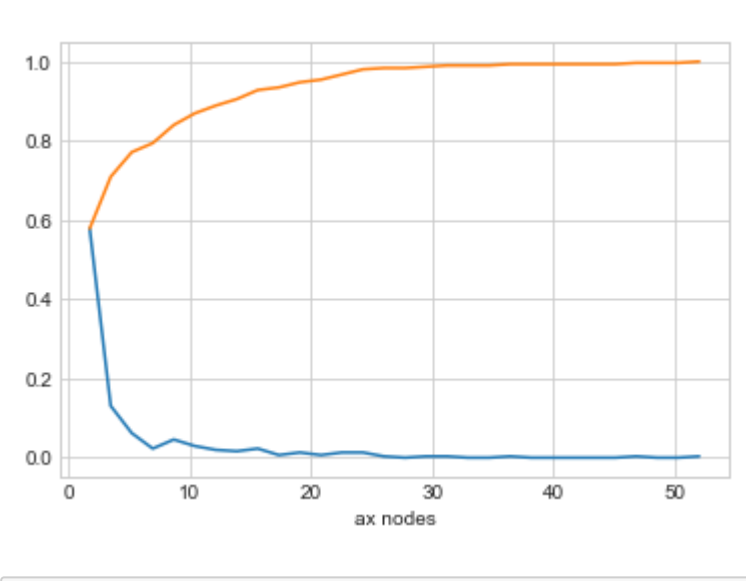
```
[0.01633987 0.03594771 0.02614379 0.0620915  0.09803922 0.05228758
 0.08169935 0.09150327 0.08169935 0.09803922 0.08496732 0.04901961
 0.06535948 0.06862745 0.01960784 0.03921569 0.01633987 0.00653595
 0.00326797 0.00326797]
```



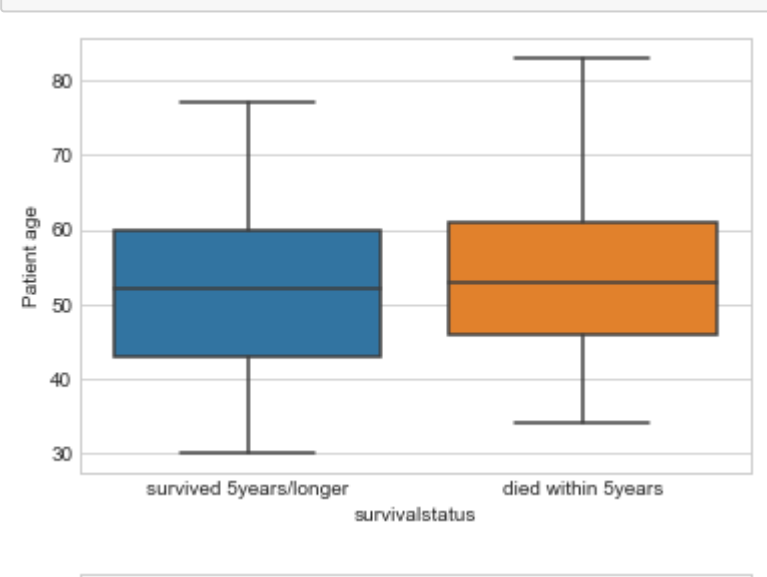

```
[0.11764706 0.08823529 0.         0.09150327 0.         0.08496732
 0.         0.0751634  0.         0.09803922 0.10130719 0.
 0.09150327 0.         0.09150327 0.         0.08169935 0.
 0.04248366 0.03594771]
```

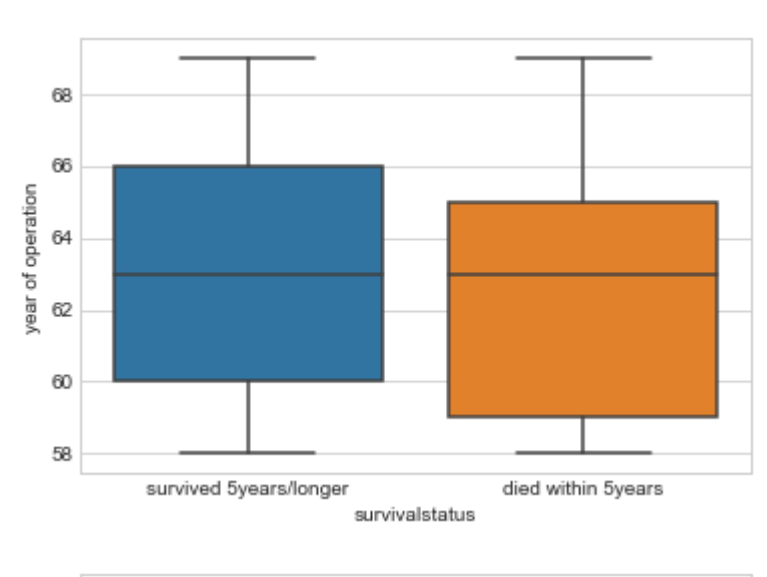


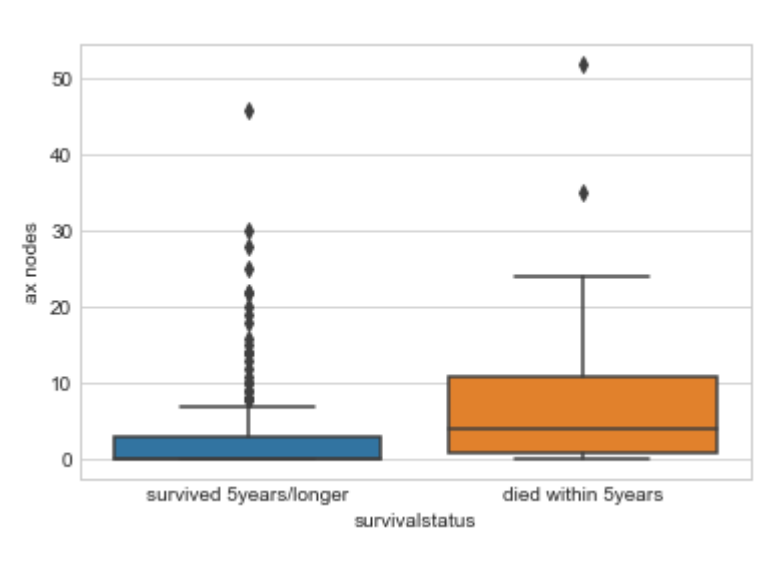

```
[0.57843137 0.13071895 0.0620915  0.02287582 0.04575163 0.02941176
 0.01960784 0.01633987 0.02287582 0.00653595 0.0130719  0.00653595
 0.0130719  0.0130719  0.00326797 0.         0.00326797 0.00326797
 0.         0.         0.00326797 0.         0.         0.
 0.         0.         0.00326797 0.         0.         0.00326797]
```



In [252]:
```
#BOX PLOT and Whiskers
sns.boxplot(x="survivalstatus",y="Patient age",data=haberman)
plt.show()
sns.boxplot(x="survivalstatus",y="year of operation",data=haberman)
plt.show()
sns.boxplot(x="survivalstatus",y="ax nodes",data=haberman)
plt.show()
```





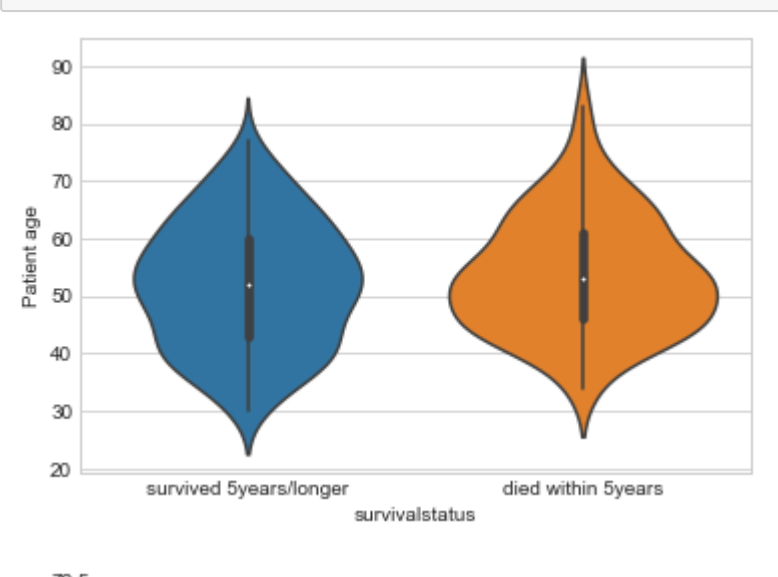


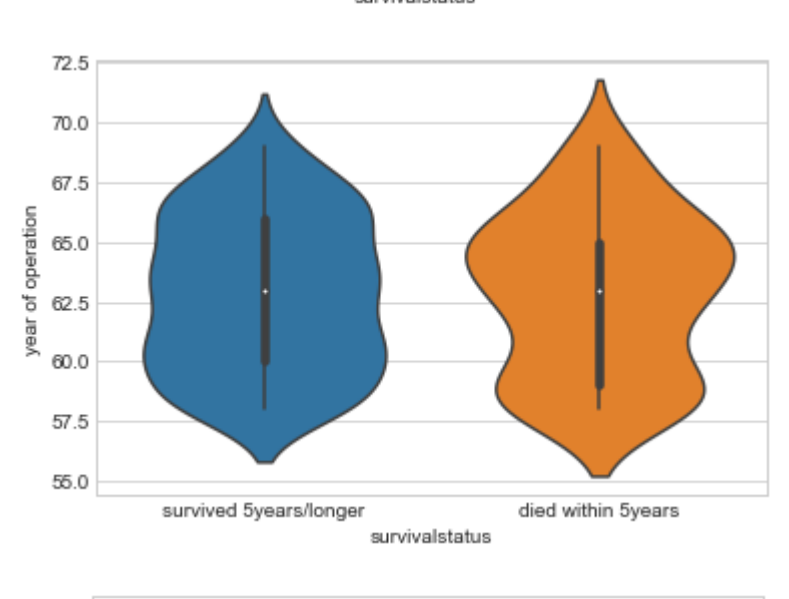
Observation-1.From box plot-[Patient Age]-we noticed that after the age 1942 , 25% of the patients observed with cancer.

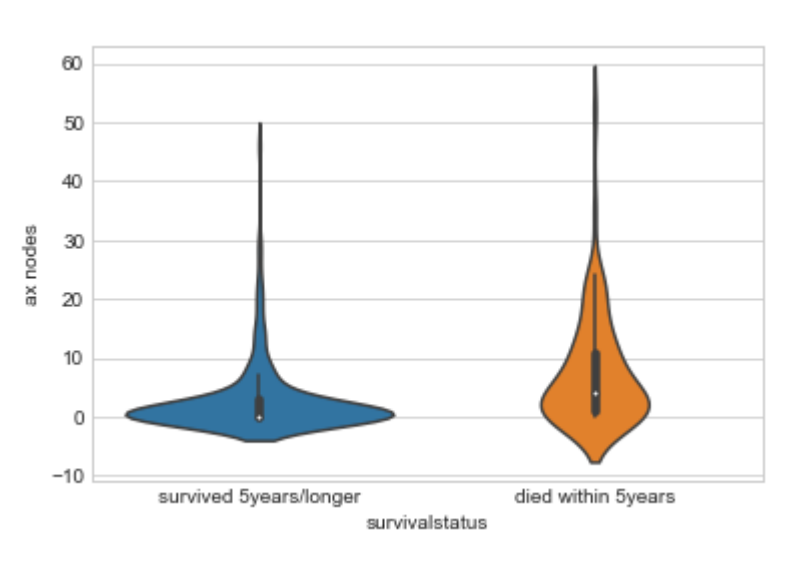
1. from box plot-[year of operation]-we noticed that after the year of 1965, 75 % of the patients survived longer than five-years as the invention of new medicines and technology in medical field.

3.from box plot-[ax nodes]- we noticed that 75 % of the people survived 5 years or longer has ax nodes is less than 5.

In [257]:
```
#violin plots
sns.violinplot(x="survivalstatus",y="Patient age",data=haberman)
plt.show()
sns.violinplot(x="survivalstatus",y="year of operation",data=haberman)
plt.show()
sns.violinplot(x="survivalstatus",y="ax nodes",data=haberman)
plt.show()
```



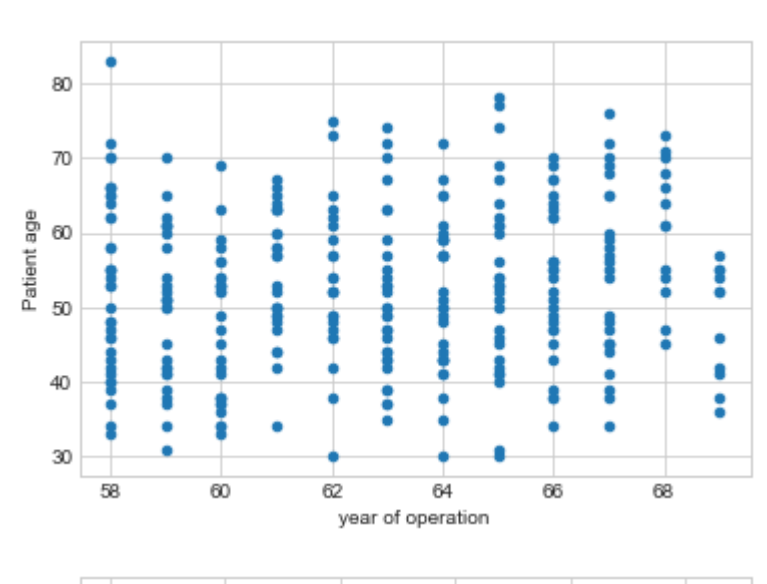


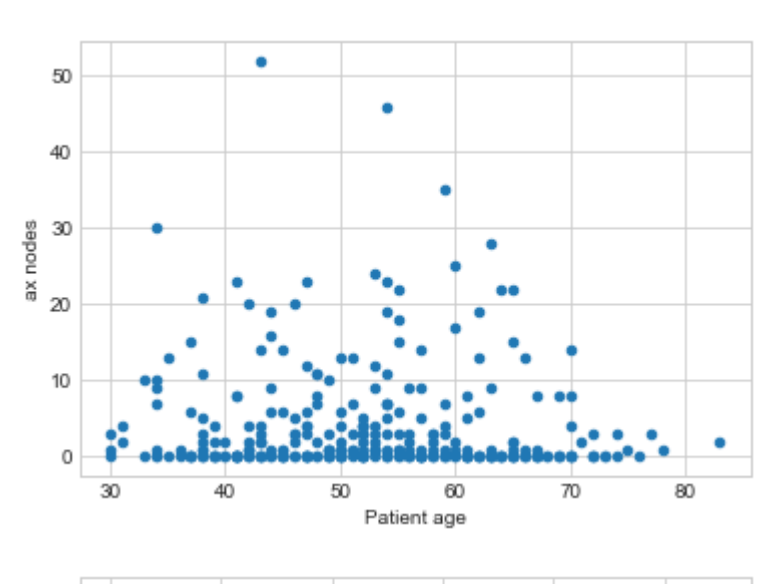


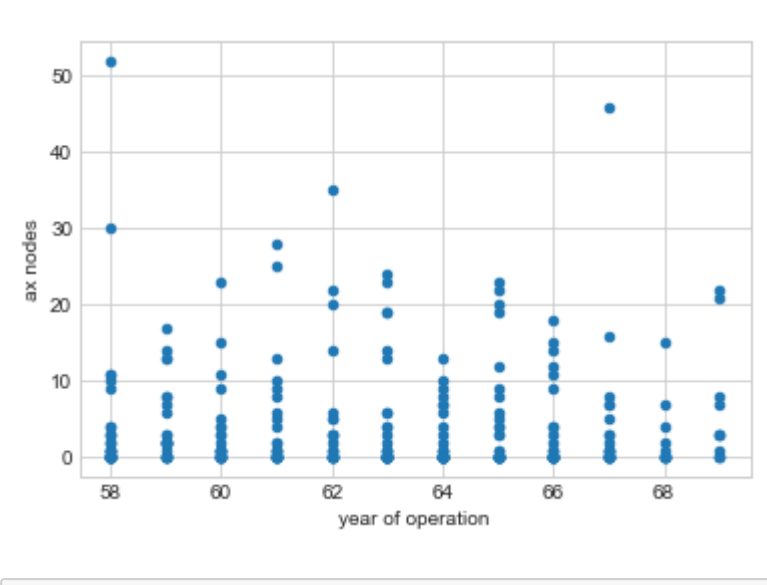
Observation-Violin plot is a combine form of PDF,Boxplot and Whiskers. By this plot according to survival status,with 75% accuracy.

In [241]:
```
#2.Bivariate Analysis
#2D scatter plot
haberman.plot(kind='Scatter',x='year of operation',y='Patient age')
print(plt.plot())
haberman.plot(kind='Scatter',x='Patient age',y='ax nodes')
print(plt.plot())
haberman.plot(kind='Scatter',x='year of operation',y='ax nodes')
print(plt.plot())
```

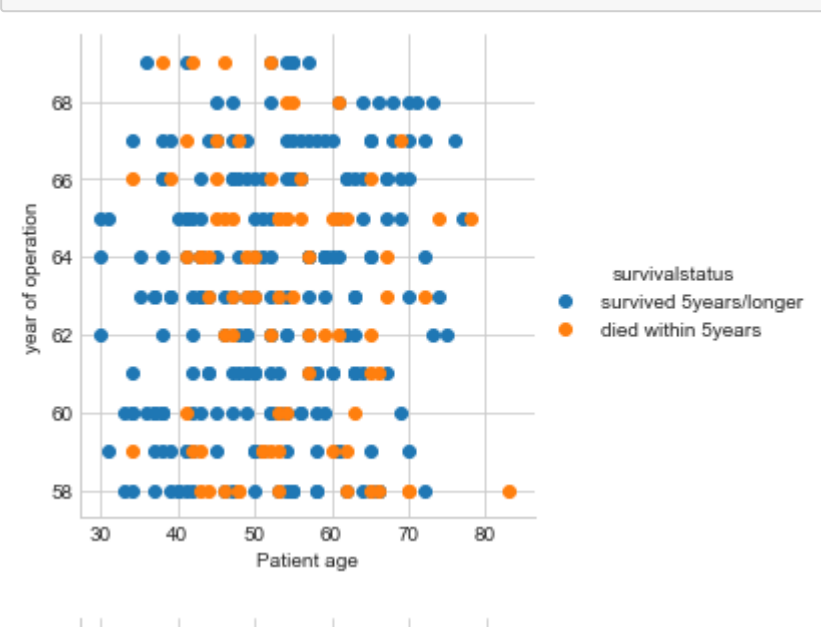
```
[]
[]
[]
```

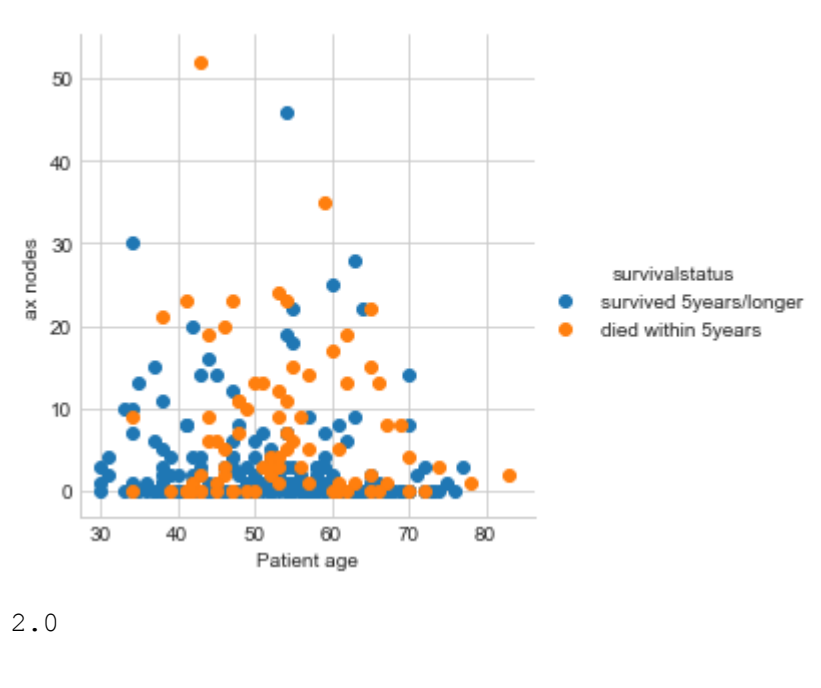






In [243]:
```
#2D scatter plot with color coding
sns.set_style("whitegrid")
sns.FacetGrid(haberman,hue="survivalstatus",size=4)\
.map(plt.scatter,"Patient age","year of operation")\
.add_legend()
sns.FacetGrid(haberman,hue="survivalstatus",size=4)\
.map(plt.scatter,"Patient age","ax nodes")\
.add_legend()
#sns.FacetGrid(haberman,hue="survivalstatus",size=4)\
#.map(plt.scatter,"year of operation","ax nodes")\
#.add_legend()
plt.show()
```



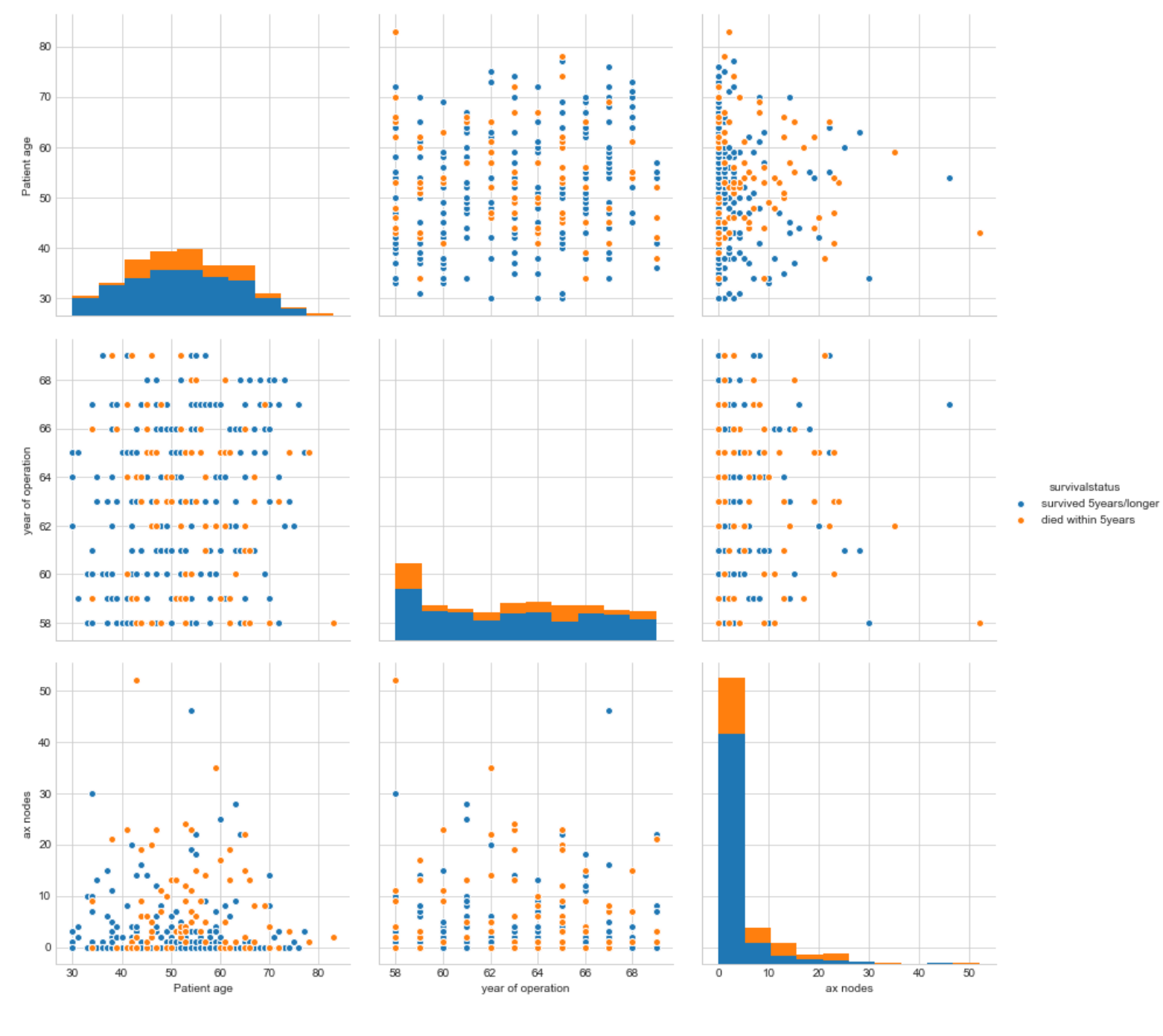


Out[243]: 2.0

Observation-1.from year of operation v/s Patient age plot we observed that at the age 42 to 65 patients has cancer. and after the year of 65 patients are surviving 5 years or longer as the technology is increasing in Medical field.

2.Patient age v/s ax nodes -Most of the patients have zero ax nodes.

In [244]:
```
#pair plot
plt.close()
sns.set_style("whitegrid")
sns.pairplot(haberman,hue="survivalstatus",size=4)
plt.show()
```

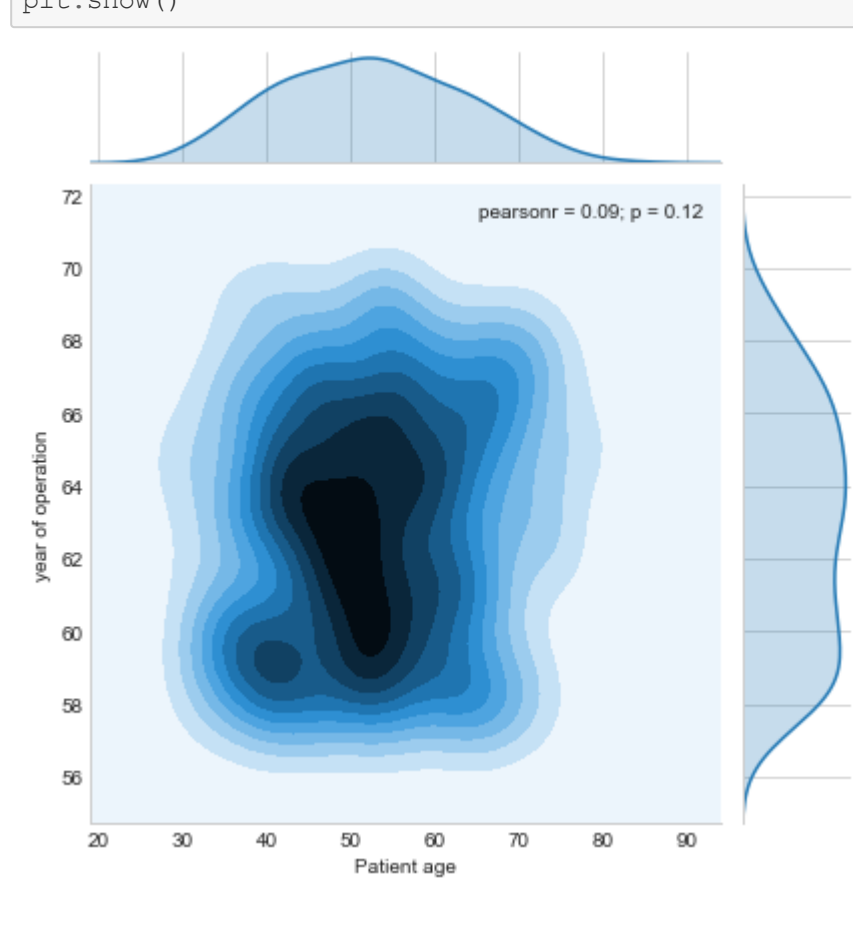


Observation-1.Year of operation and Patient age is the most useful features to identify that at the age of 42-65 Patients has cancer.

2.after the year of 65 patients are surviving 5 years or longer as the technology is increasing in Medical field.

3.we can use if else conition to build a simple model to ientify the survivalstatus and the age of Patients who suffered from cancer.

In [262]:
```
#3.Multivariate Probability density,Contour Plot
sns.jointplot(x="Patient age",y="year of operation",data=haberman,kind='kde')
plt.show()
```



Observation- From contour plot we can say that max number of patients undergone operation during the year 59 to 64 period & b/w ages 42 to 60.

CONCLUSION- The main conclusion of haberman dataset we observed that- 1.In haberman dataframe 306 rows,and 4 columns is present. 2.It is Imbalanced data set. 3.columns names are-['Patient age','year of operation','ax nodes','survivalstatus']. 4.survivalstatus has 2 different class of Patients- among all the patients 225 number of patients can survived 5 years or longer & 81 Patients died within 5 years. 5.a)From box plot-[Patient Age]-we noticed that after the age 1942 , 25% of the patients observed with cancer.

b)from box plot-[year of operation]-we noticed that after the year of 1965, 75 % of the patients survived longer than five-years as the invention of new medicines and technology in medical field.

c)from box plot-[ax nodes]- we noticed that 75 % of the people survived 5 years or longer has ax nodes is less than 5.

6.[from Pair-plot]-1.Year of operation and Patient age is the most useful features to identify that at the age of 42-65 Patients has cancer.

2.after the year of 65 patients are surviving 5 years or longer as the technology is increasing in Medical field.

3.we can use if else condition to build a simple model to identify the survivalstatus and the age of Patients who suffered from cancer.