In [1]:

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

In [2]:

```
pwd
```

Out[2]:

```
'C:\\Users\\HP\\EDA'
```

In [3]:

```
df=pd.read_csv(r'C:\Users\HP\Downloads\EDA\haberman.csv' )
```

In [4]:

```
df.head()
```

Out[4]:

	patient_age	year_of_operation	positive_axillary_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

In [5]:

```
df.shape
```

Out[5]:

```
(306, 4)
```

In [6]:

```
df.describe
```

Out[6]:

```
<bound method NDFrame.describe of      patient_age  year_of_operation  positive_axillary_nodes  sur
vival_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]>
```

no missing values

In [7]:

```
df['survival_status'] = df['survival_status'].map({1:'survived', 2:'dead'})
```

In [8]:

```
df.tail()
```

Out[8]:

	patient_age	year_of_operation	positive_axillary_nodes	survival_status
301	75	62	1	survived
302	76	67	0	survived
303	77	65	3	survived
304	78	65	1	dead
305	83	58	2	dead

In [9]:

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

Out[9]:

```
survived    225
dead         81
Name: survival_status, dtype: int64
```

imbalanced data set

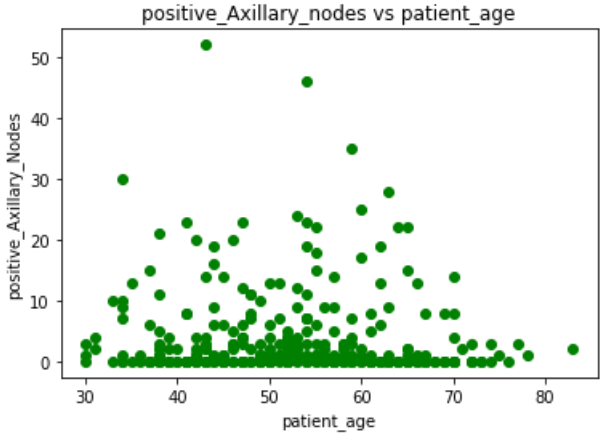
In [ ]:

```
#scatterplot
plt.scatter(df['patient_age'],df['year_of_operation'], color = 'g')
plt.xlabel('patient_age')
plt.ylabel('year_of_operation')
plt.title('year_of_opeation vs patient_age')
plt.show()
```

map doenst clear shows data points but opeartions doen between ages 40 to 68

In [11]:

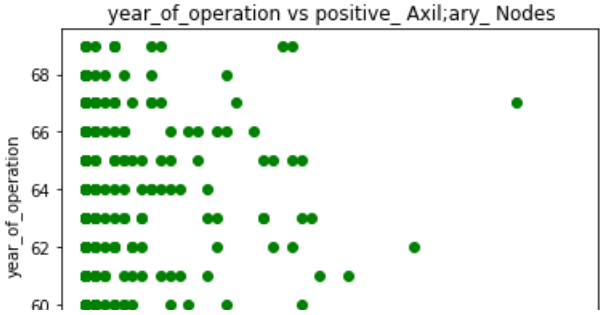
```
#scatter plot
plt.scatter(df['patient_age'],df['positive_axillary_nodes'], color = 'g')
plt.xlabel('patient_age')
plt.ylabel('positive_Axillary_Nodes')
plt.title('positive_Axillary_nodes vs patient_age')
plt.show()
```

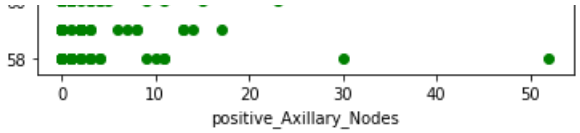


concentration at axillary node 0

In [13]:

```
#scatter plot
plt.scatter(df['positive_axillary_nodes'], df['year_of_operation'], c = 'g')
plt.xlabel('positive_Axillary_Nodes')
plt.ylabel('year_of_operation')
plt.title(' year_of_operation vs positive_ Axil;ary_ Nodes')
plt.show()
```

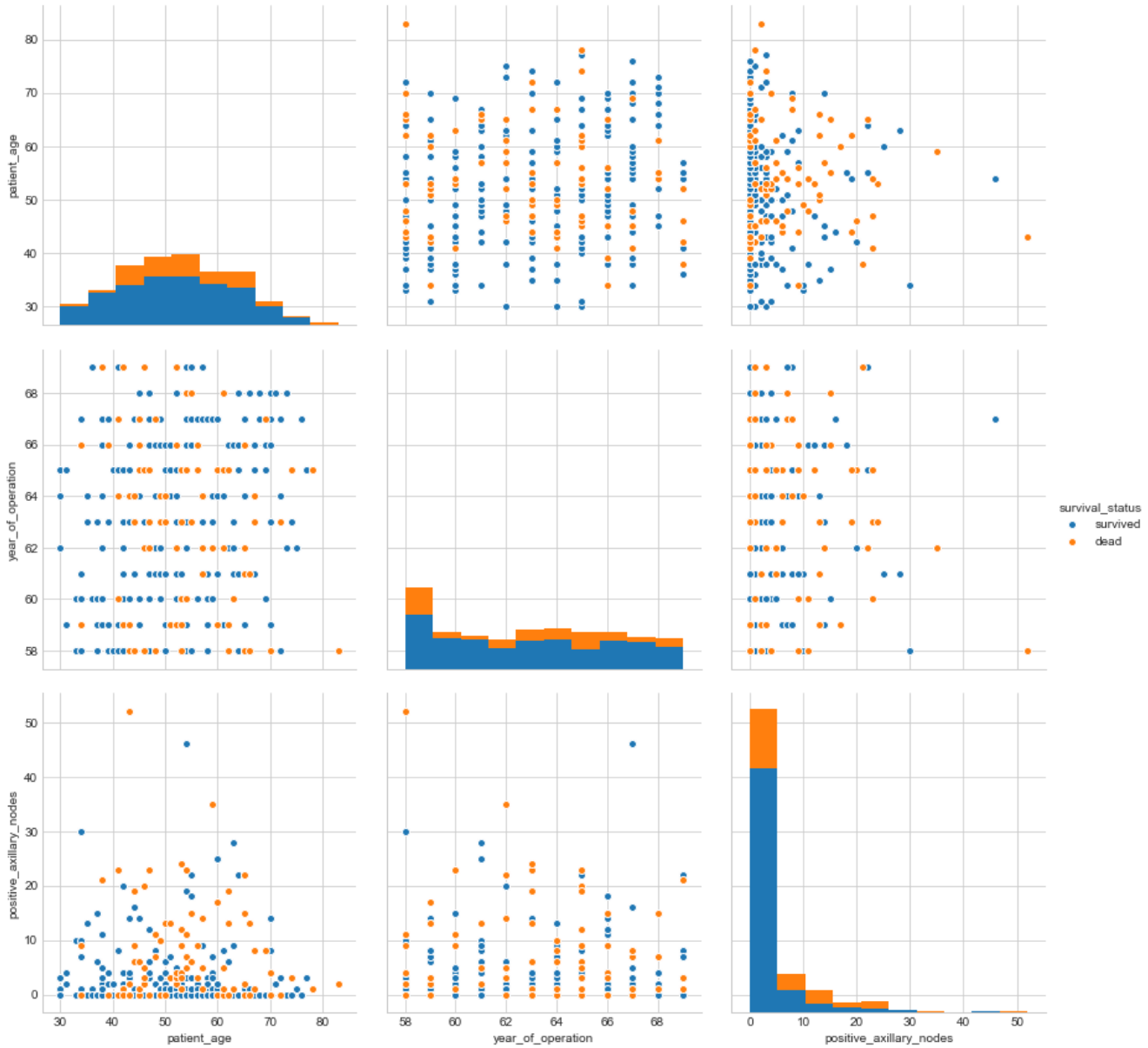




most operations done between 1960 and 1966

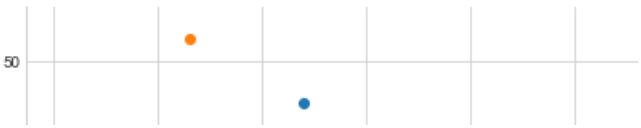
In [12]:

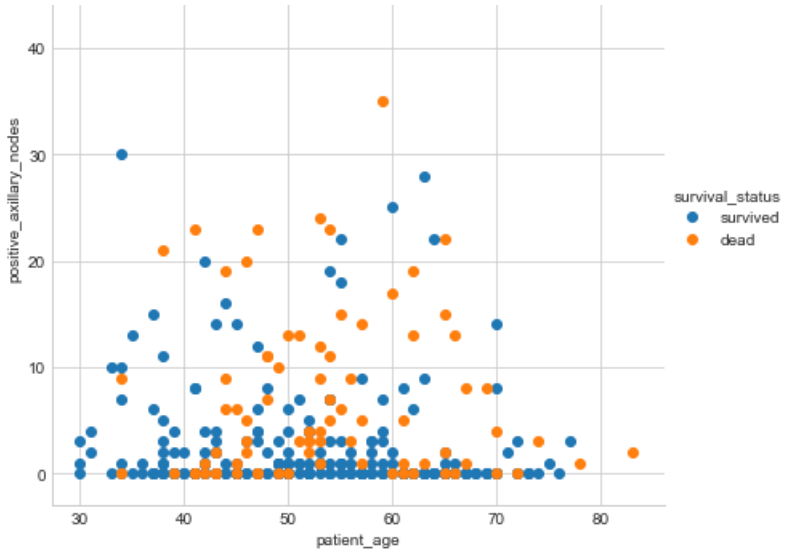
```
#pairplot
plt.close();
sns.set_style('whitegrid');
sns.pairplot(df, hue = 'survival_status', size = 4)
plt.show()
```



In [15]:

```
#satterplot
sns.set_style('whitegrid');
sns.FacetGrid(df, hue = 'survival_status', size = 6)\
   .map(plt.scatter, 'patient_age', 'positive_axillary_nodes')\
   .add_legend();
plt.show();
```

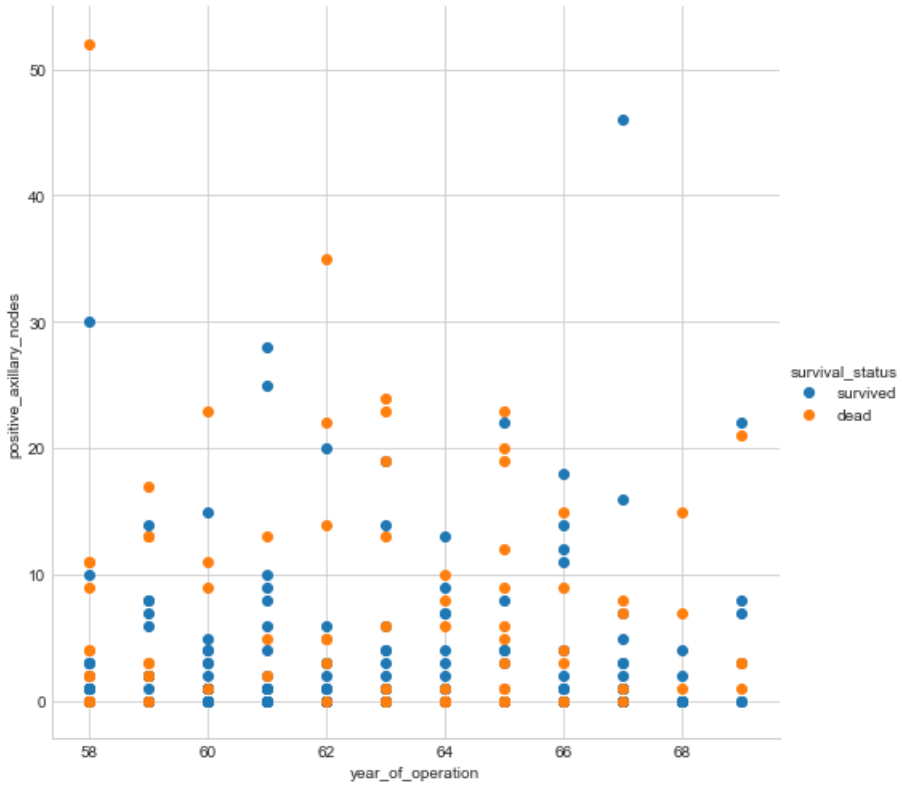




patient near to auxillary node will survive more and patient age above 50 and auxillary node above 10 will die more

In [13]:

```
#scatter plot
sns.set_style('whitegrid');
sns.FacetGrid(df, hue='survival_status', size = 7) \
    .map(plt.scatter, 'year_of_operation', 'positive_axillary_nodes') \
    .add_legend();
plt.show()
```



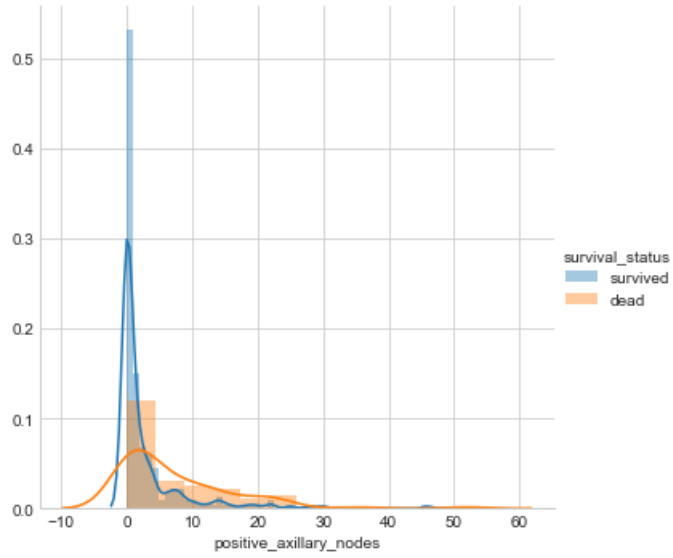
no clear information here

In [14]:

```
#Distribution plot
sns.FacetGrid(df, hue='survival_status', size = 5) \
    .map(sns.distplot, 'positive_axillary_nodes') \
    .add_legend();
plt.show();
```

```
C:\ProgramData\Anaconda3\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:\ProgramData\Anaconda3\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 "
```

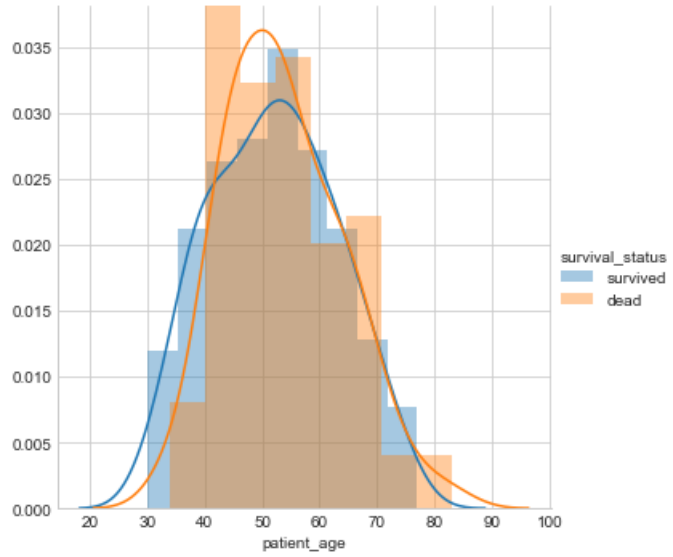


patients having 0 auxillary nodes will surive more

In [15]:

```
#Distribution plot
sns.FacetGrid(df, hue='survival_status', size = 5) \
    .map(sns.distplot, 'patient_age') \
    .add_legend();
plt.show();
```

```
C:\ProgramData\Anaconda3\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:\ProgramData\Anaconda3\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 "
```



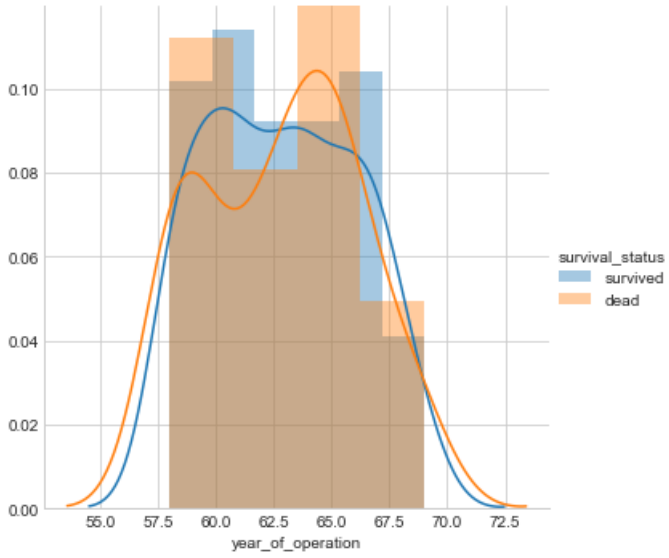
age between 40 to 60 likely to die and age less than 40 have survive more

In [16]:

```
#Distribution plot
sns.FacetGrid(df, hue='survival_status', size = 5) \
    .map(sns.distplot, 'year_of_operation') \
    .add_legend();
```

```
plt.show();
```

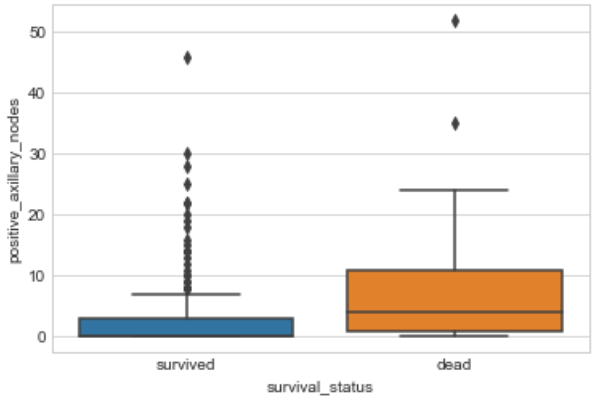
```
C:\ProgramData\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The
'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\ProgramData\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: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



large patients whose operation doen between 60 and 65

In [17]:

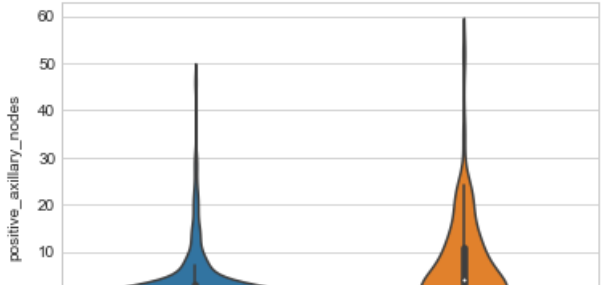
```
#Boxplot
sns.boxplot(x='survival_status', y = 'positive_axillary_nodes', data=df)
plt.show()
```



more number of auxillary nodes more likely to die

In [18]:

```
#violinplot
sns.violinplot(x='survival_status', y='positive_axillary_nodes', data = df, size = 9)
plt.show()
```



patients having auxillary nodes at 0 will surive more and towards or 1 will die more

patients at auxillary node 0 and age less then 40 and year of opearation done between 1960 and 1965 will survive

final conclusion

1.From this Dataset we can say that the majority of operations are performed on people age range between 38 and 68, where most of the points plotted on scatter plot (year_of_Operation vs patient_age).

2.We can see that there is quite good concentration of data point When positive_axillary_node is 0.

3.we can see that most operations done between 1960 and 1966.

4.Here with this scatter plot we get insight that patients with 0 axil nodes are more likely to survive .

5.Patients who are older than 50 and have axil nodes greater than 10 are more likely to die.

6.we can observe patients having 0 auxillary nodes will surive more . 7.we conclude that age between 40 to 60 likely to die and age less than 40 have survive more.

8.we can see more number of auxillary nodes more likely to die and patients having auxillary nodes at 0 will surive more and towards or 1 will die more.