In [1]:

```
##.......IMPORTING LIBRARY.....
import pandas as PD
import seaborn as SNS
import matplotlib.pyplot as PLT
import numpy as NP
```

In [2]:

```
#...READING...CSV FILE USING PANDAS....
DF = PD.read_csv('haberman.csv')
```

In [3]:

```
DF.head(4)
```

Out[3]:

	30	64	1	1.1
0	30	62	3	1
1	30	65	0	1
2	31	59	2	1
3	31	65	4	1

**...ASSINGNING COLUMN NAME AS : 1.AGE 2.YEAR_OF_OPERATION 3.NUMBER_OF_POSITIVE_NODES 4.SURVIVAL_STATUS

In [4]:

```
DF.columns = ['AGE','YEAR_OF_OPERATION','NUMBER_OF_POSITIVE_NODES','SURVIVAL_STATUS']
```

In [5]:

```
#...first 4 rows of our data
DF.head(4)
```

Out[5]:

	AGE	YEAR_OF_OPERATION	NUMBER_OF_POSITIVE_NODES	SURVIVAL_STATUS
0	30	62	3	1
1	30	65	0	1
2	31	59	2	1
3	31	65	4	1

**...Replacing "1" = SURVIVED and "2" = NOT_SURVIVED

In [6]:

```
DF['SURVIVAL_STATUS'] = DF['SURVIVAL_STATUS'].map({1:'SURVIVED',2:'NOT_SURVIVED'})
```

In [7]:

```
DF.head(4)
```

Out[7]:

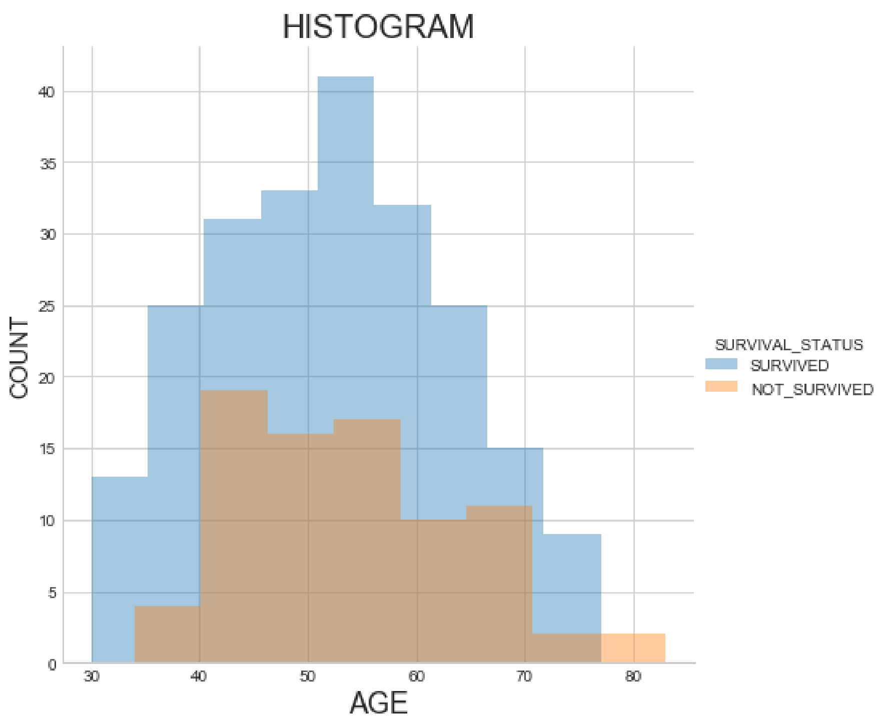
	AGE	YEAR_OF_OPERATION	NUMBER_OF_POSITIVE_NODES	SURVIVAL_STATUS
0	30	62	3	SURVIVED
1	30	65	0	SURVIVED
2	31	59	2	SURVIVED
3	31	65	4	SURVIVED

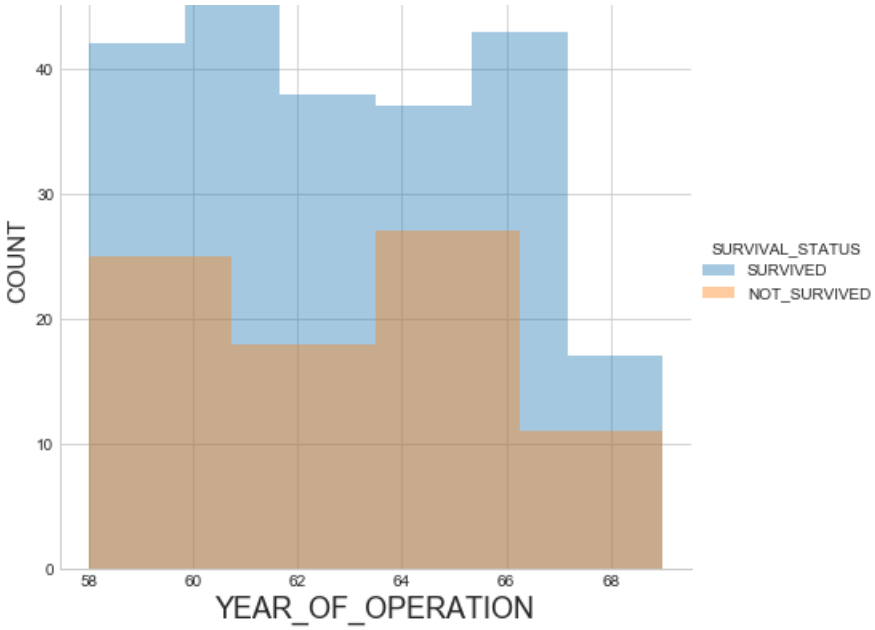
Here we did univariate analysis...

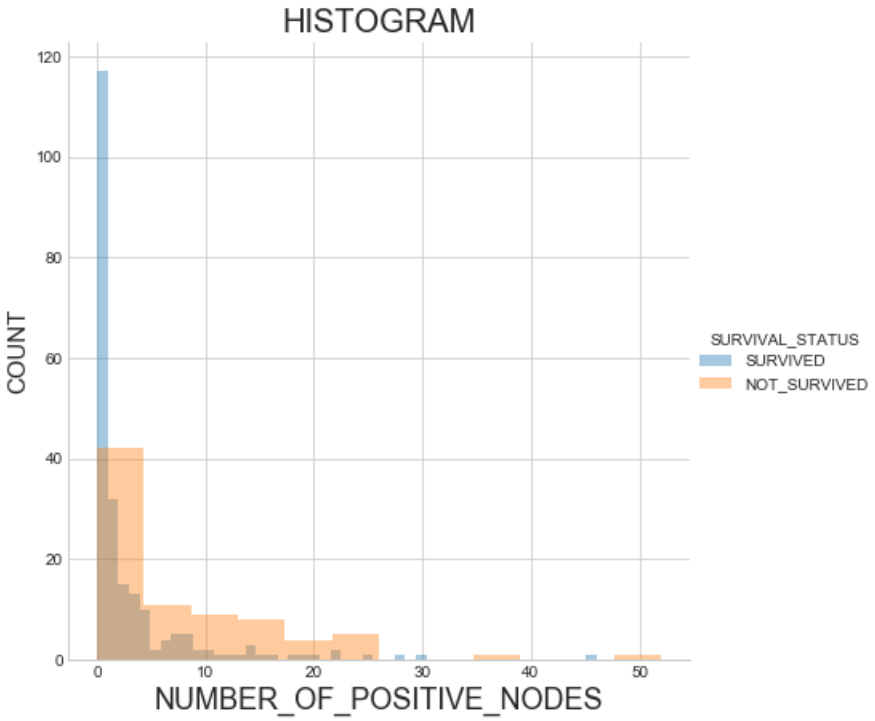
In [8]:

```
for i in range(0,len(DF.columns)-1):
    SNS.set_style("whitegrid")
    OBJ = SNS.FacetGrid(DF, hue= 'SURVIVAL_STATUS',size = 6)
    OBJ.map(SNS.distplot,DF.columns[i],kde =False).add_legend()
    #OBJ.add_legend()
    PLT.xlabel(DF.columns[i],fontsize = 18)
    PLT.ylabel('COUNT',fontsize = 15)
    PLT.title('HISTOGRAM',fontsize = 20)
```

```
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'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.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 "
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 "
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 "
```







Inference:

1. 40 patient lie in between age group of 50 to 55 has Survived.
2. 30 patient lie in between age group of 45 to 50 and 55 to 60 has survived.
3. There are almost 20 patient who lie in age group of 40 to 45 not able to survived.
4. In year span of 1960 to 1962 survived rate of patient is high.

If we Increase the bin size we can infer more exact information.
In following kernel we have increased bin size (40) and we can clearly extract more insight in plot of Year of opertion as shown below:

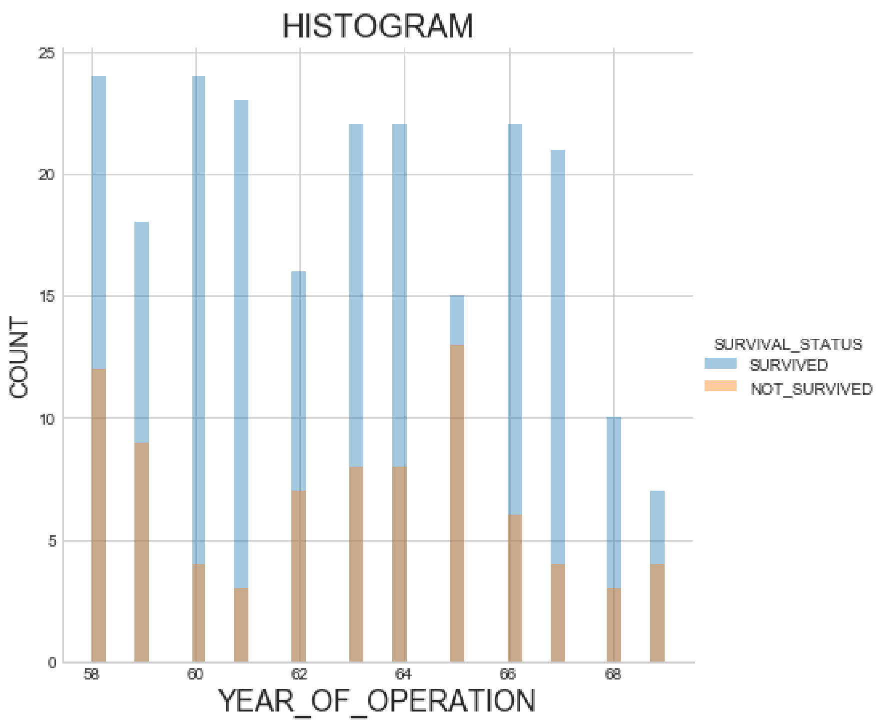
In [9]:

```
SNS.set_style("whitegrid")
OBJ = SNS.FacetGrid(DF, hue= 'SURVIVAL_STATUS',size = 6)
OBJ.map(SNS.distplot,DF.columns[1],kde =False,bins = 40).add_legend()
    #OBJ.add_legend()
PLT.xlabel(DF.columns[1],fontsize = 18)
PLT.ylabel('COUNT',fontsize = 15)
PLT.title('HISTOGRAM',fontsize = 20)
```

```
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 "
```

Out[9]:

```
Text(0.5,1,'HISTOGRAM')
```



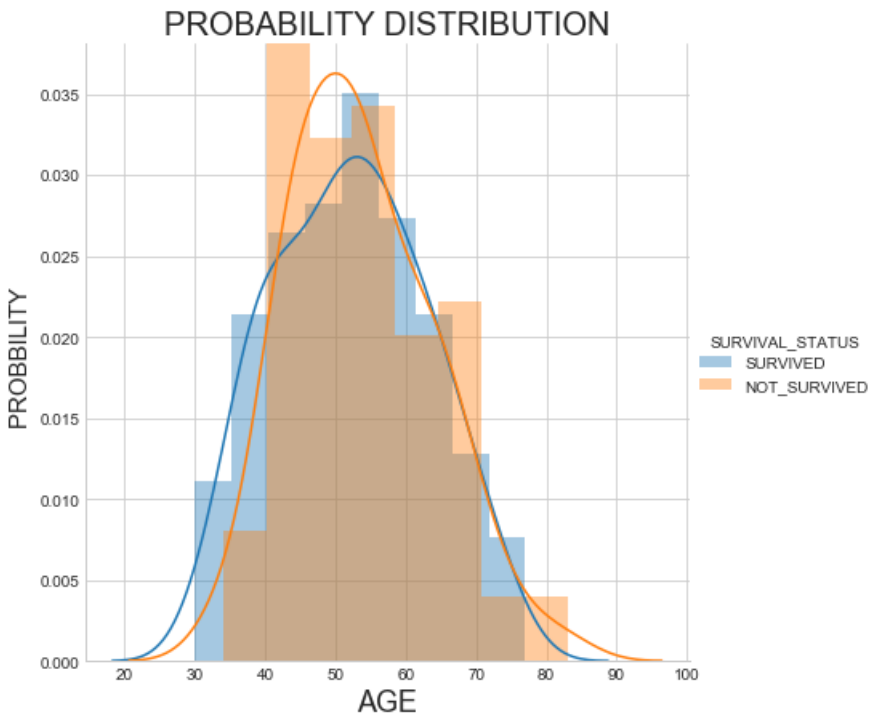
**AS we increase bin size to 40 we got more insight of data. 1.In year 1961 almost 75 % of people has survived. 2.In year 1965 almost 85% of people not able to survived.

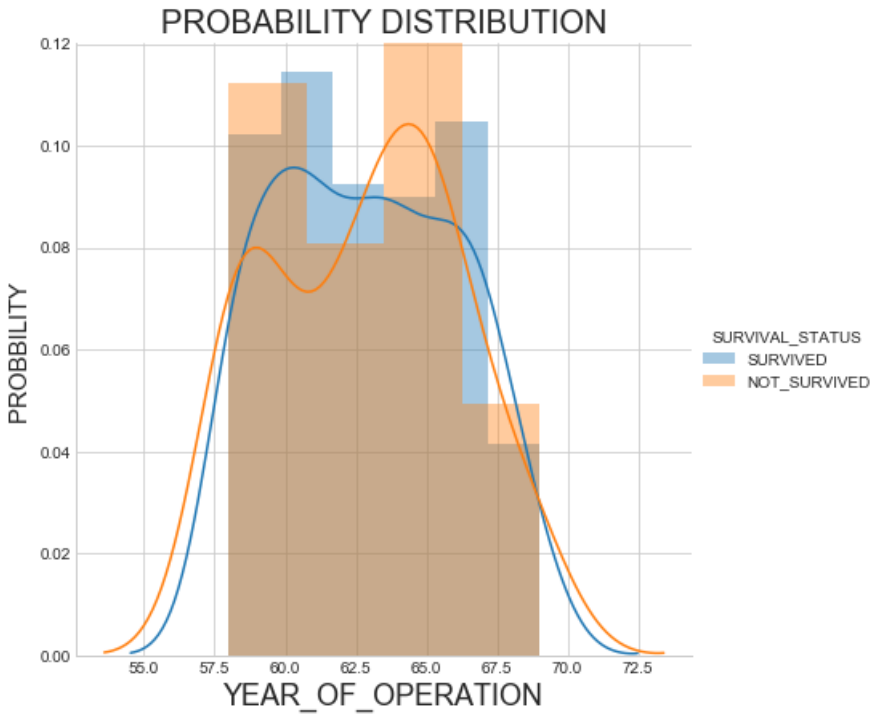
** Similar Inferences can be extract if we plot Kernel density estimation of above plots.

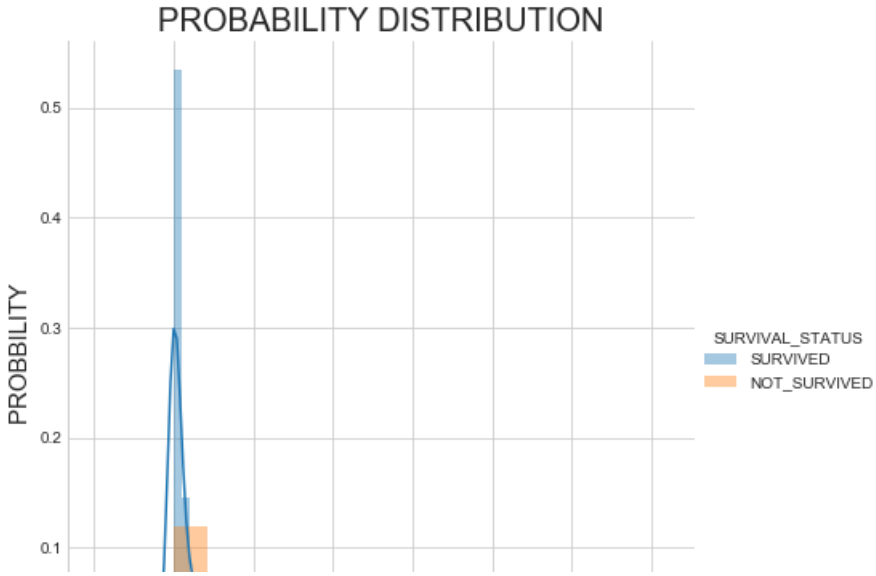
In [10]:

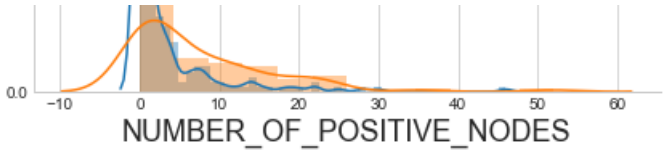
```python
for i in range(0,len(DF.columns)-1):
    SNS.set_style("whitegrid")
    OBJ = SNS.FacetGrid(DF, hue= 'SURVIVAL_STATUS',size = 6)
    OBJ.map(SNS.distplot,DF.columns[i],kde =True,).add_legend()
    #OBJ.add_legend()
    PLT.xlabel(DF.columns[i],fontsize = 18)
    PLT.ylabel('PROBBILITY ',fontsize = 15)
    PLT.title('PROBABILITY DISTRIBUTION',fontsize = 20)
```

```
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 "
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 "
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 "
```


PROBABILITY DISTRIBUTION
PROBBILITY
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
20
30
40
50
60
70
80
90
100
AGE
SURVIVAL_STATUS
SURVIVED
NOT_SURVIVED


PROBABILITY DISTRIBUTION
PROBBILITY
0.12
0.10
0.08
0.06
0.04
0.02
0.00
55.0
57.5
60.0
62.5
65.0
67.5
70.0
72.5
YEAR_OF_OPERATION
SURVIVAL_STATUS
SURVIVED
NOT_SURVIVED


PROBABILITY DISTRIBUTION
PROBBILITY
0.5
0.4
0.3
0.2
0.1
SURVIVAL_STATUS
SURVIVED
NOT_SURVIVED



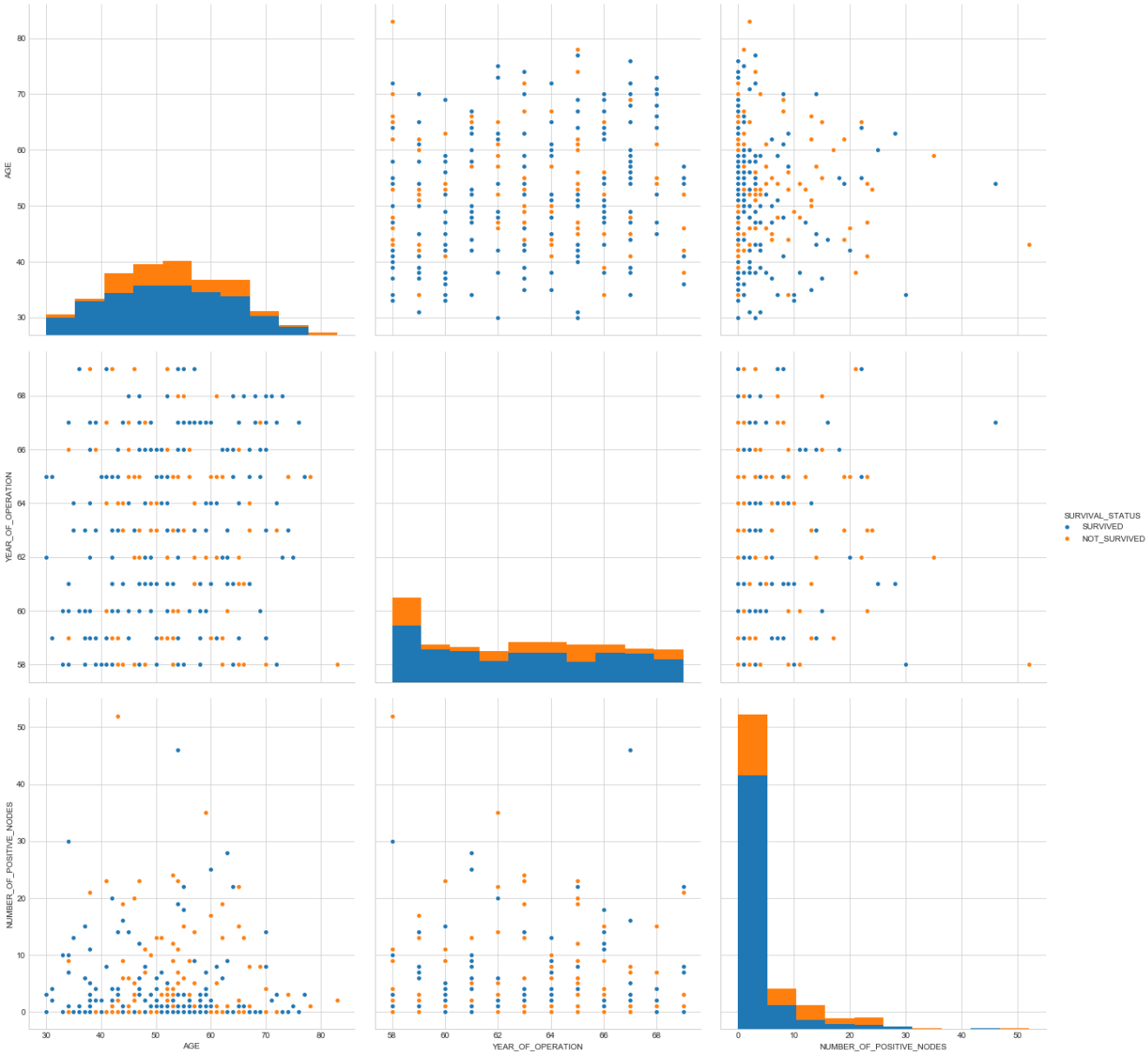
2D analysis using Pairplot

In [11]:

```
SNS.pairplot(DF, hue='SURVIVAL_STATUS', size=6)
```

Out[11]:

```
<seaborn.axisgrid.PairGrid at 0x177195f2d30>
```



Inference:

** AGE vs Year Of Operation: 1.After 1968 all patient age were less than 60. 2.In year 1965 we have so many failure operation.

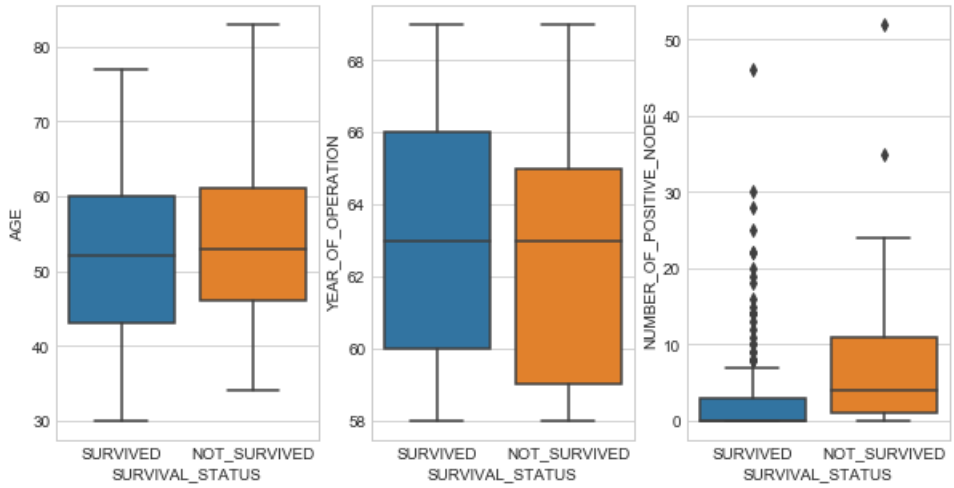
**AGE vs Number of positive Nodes: 1.Most of the patient have number of positive nodes in between 0 to 10.If patient age is less than 70 there is more than 50 % of chance that patient will survived.

**Number of Positive Nodes Vs Year Of Operation: 1.only in year 1958 there was a case in which number of positive nodes is greater than 50.

BOX Plot and Violin Plot

In [12]:

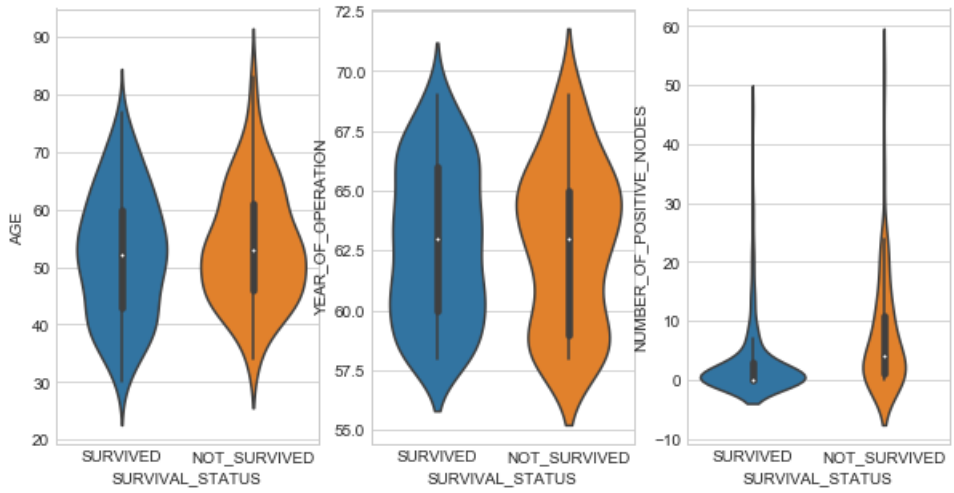
```
FIG ,AX_ARR = PLT.subplots(1,3,figsize = (10,5))
for i in range(0,len(DF.columns)-1):
    SNS.boxplot(x = 'SURVIVAL_STATUS',y = DF.columns[i],data = DF,ax = AX_ARR[i])
```



**Whatever infernce we made out by looking CDF can be easily made by looking BOX PLOT. 1.In first plot 25^th percentile value is approx equal to 45 it state that almost 75%percent of people having age is greater than 45.

In [13]:

```
FIG ,AX_ARR = PLT.subplots(1,3,figsize = (10,5))
for i in range(0,len(DF.columns)-1):
    SNS.violinplot(x = 'SURVIVAL_STATUS',y = DF.columns[i],data = DF,ax = AX_ARR[i])
```



**WE can infer so many things like median,PDF,all observation from box plot all those thing can be conclude from violin plot

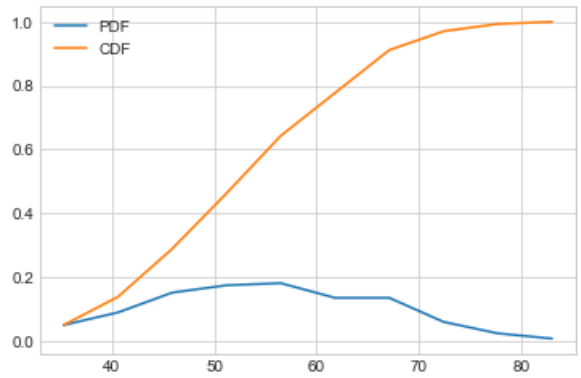
PDF and CDF :

In [14]:

```
COUNT,B = NP.histogram(DF.AGE,bins = 10,density = True)
PDF = COUNT/sum(COUNT)
CDF = NP.cumsum(PDF)
PLT.plot(B[1:],PDF)
PLT.plot(B[1:],CDF)
PLT.legend(['PDF','CDF'])
```

Out[14]:

```
<matplotlib.legend.Legend at 0x1771c11ee80>
```



**CDF will tell us about percentile value as we can see from the above CDF plot 50^th percentile value lie is approx equal to 50.

In [15]:

```
#...VErifying above statement.
print("*****___INFERENCE FROM CDF_____*****")
print("10% of patient have age greater than {}".format(NP.percentile(DF['AGE'],90)))
print("40% of patient have age less than {}".format(NP.percentile(DF["AGE"],40)))
```

```
*****___INFERENCE FROM CDF_____*****
10% of patient have age greater than 67.0
40% of patient have age less than 49.0
```

- We can verify same result using BOX PLOT.

Inference:

1.Almost 60% area under PDF cover by age group of 45 to 65 2.Now we Plotted seperate CDF and PDF of people who Survived and Who didn't.

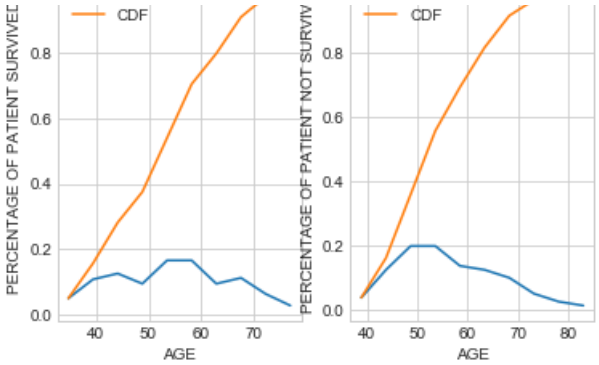
In [16]:

```
PLT.subplot(121)
DF_SURVIVED = DF[DF['SURVIVAL_STATUS']=='SURVIVED']
COUNT,B = NP.histogram(DF_SURVIVED.AGE,bins  = 10,density = True)
PDF = COUNT/sum(COUNT)
CDF = NP.cumsum(PDF)
PLT.plot(B[1:],PDF)
PLT.plot(B[1:],CDF)
PLT.xlabel('AGE')
PLT.ylabel('PERCENTAGE OF PATIENT SURVIVED')
PLT.legend(['PDF','CDF'])

PLT.subplot(122)
DF_NOT_SURVIVED = DF[DF['SURVIVAL_STATUS']=='NOT_SURVIVED']
COUNT,B = NP.histogram(DF_NOT_SURVIVED.AGE,bins  = 10,density = True)
PDF = COUNT/sum(COUNT)
CDF = NP.cumsum(PDF)
PLT.plot(B[1:],PDF)
PLT.plot(B[1:],CDF)
PLT.xlabel('AGE')
PLT.ylabel('PERCENTAGE OF PATIENT NOT SURVIVED')
PLT.legend(['PDF','CDF'])

print("***___INFERNCE_FROM_CDF__***")
print("5 % of people whose having age greater than {} is
survived".format(NP.percentile(DF_SURVIVED["AGE"],95)))
```

```
***___INFERNCE_FROM_CDF__***
5 % of people whose having age greater than 70.0 is survived
```



Inference:

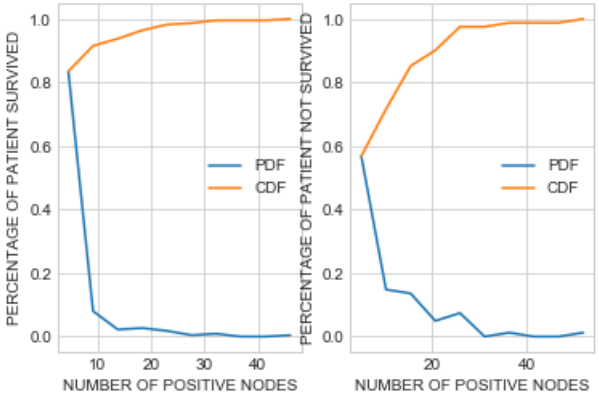
1. People who survived are mostly in lie in age group of 50 to 60.(because are under those part of distribution is high as compare to other age group.
2. People who is older than 70 year has very less chance to survive.

In [17]:

```
PLT.subplot(121)
DF_SURVIVED = DF[DF['SURVIVAL_STATUS']=='SURVIVED']
COUNT,B = NP.histogram(DF_SURVIVED.NUMBER_OF_POSITIVE_NODES,bins  = 10,density = True)
PDF = COUNT/sum(COUNT)
CDF = NP.cumsum(PDF)
PLT.plot(B[1:],PDF)
PLT.plot(B[1:],CDF)
PLT.xlabel('NUMBER OF POSITIVE NODES')
PLT.ylabel('PERCENTAGE OF PATIENT SURVIVED')
PLT.legend(['PDF','CDF'])
PLT.subplot(122)
DF_NOT_SURVIVED = DF[DF['SURVIVAL_STATUS']=='NOT_SURVIVED']
COUNT,B = NP.histogram(DF_NOT_SURVIVED.NUMBER_OF_POSITIVE_NODES,bins  = 10,density = True)
PDF = COUNT/sum(COUNT)
CDF = NP.cumsum(PDF)
PLT.plot(B[1:],PDF)
PLT.plot(B[1:],CDF)
PLT.xlabel('NUMBER OF POSITIVE NODES')
PLT.ylabel('PERCENTAGE OF PATIENT NOT SURVIVED')
PLT.legend(['PDF','CDF'])


print("***___INFERNCE_FROM_CDF__***")
print("2 % of people whose having positive nodes greater than {} is survived".format(NP.percentile
(DF_SURVIVED["NUMBER_OF_POSITIVE_NODES"],98)))
print("10 % of people whose having positive nodes greater than {} is Not
survived".format(NP.percentile(DF_NOT_SURVIVED["NUMBER_OF_POSITIVE_NODES"],90)))
```

```
***___INFERNCE_FROM_CDF__***
2 % of people whose having positive nodes greater than 22.0 is survived
10 % of people whose having positive nodes greater than 20.0 is Not survived
```



Inference:

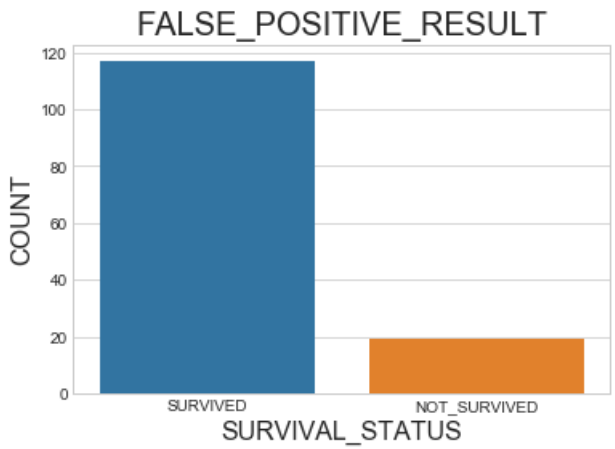
1. People who survived has number of positive nodes less than 15(those people has high rate of survive)
2. People having number of positive nodes greater than 40 has very less chance to survive.

In [18]:

```
DF_ZERO_NODES = DF[DF['NUMBER_OF_POSITIVE_NODES']==0]
SNS.countplot(DF_ZERO_NODES['SURVIVAL_STATUS'])
PLT.xlabel('SURVIVAL_STATUS',fontsize = 16)
PLT.ylabel('COUNT',fontsize = 16)
PLT.title('FALSE_POSITIVE_RESULT',fontsize = 20)
```

Out[18]:

```
Text(0.5,1,'FALSE_POSITIVE_RESULT')
```



In [19]:

```
DF_ZERO_NODES.head()
```

Out[19]:

	AGE	YEAR_OF_OPERATION	NUMBER_OF_POSITIVE_NODES	SURVIVAL_STATUS
1	30	65	0	SURVIVED
5	33	60	0	SURVIVED
6	34	59	0	NOT_SURVIVED
12	34	60	0	SURVIVED
14	35	63	0	SURVIVED

False Positive Result:

**In above kernel we observe that there was almost 135 patient in which no positive Auxillary Nodes were present ,but still they gone through operation and some of them are not able too survive.

1. Out of 135 patient almost 20 patient are there who died with in span of 5 years.
2. Almost 110 patient survived.

Inference:

**(Although this inference can't be used for our main objective which is classification,but from this analysis we can found out failure cases ,by arising question that why does category of patient who's having negative test(No Auxillary Nodes) are failed to survived and why Operation being performed on them?

1. From above kernel we know that there are some patients who do not have positive auxillary nodes but still gone through operation,now we want to see In which year this case happened and what are the age of person who did not survived.

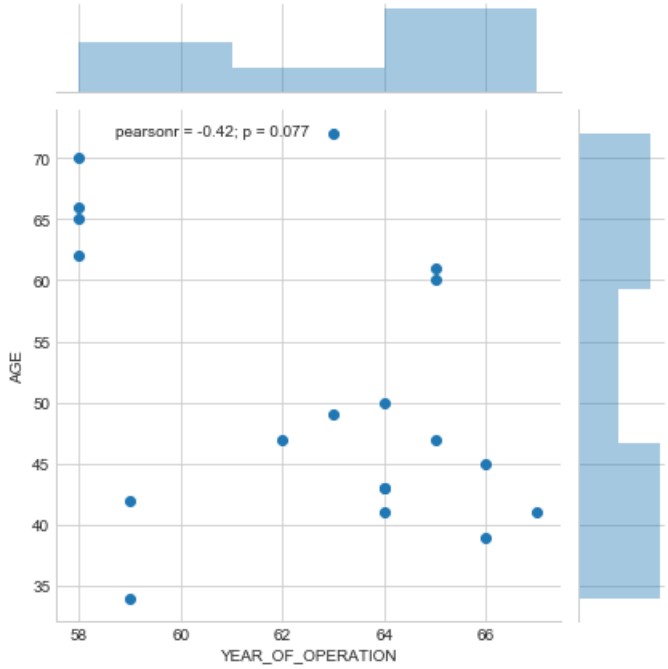
In [20]:

```
SNS.jointplot(x='YEAR_OF_OPERATION',y = 'AGE',data = DF_ZERO_NODES[DF_ZERO_NODES['SURVIVAL_STATUS']
== 'NOT_SURVIVED'])
```

```
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 "
```

Out[20]:

```
<seaborn.axisgrid.JointGrid at 0x1771c32cf98>
```



** In year 1958 there are 4 people having negative test ,but still not survived.

**Refrence: https://www.kaggle.com/ranasinghiitkgp/haberman-cancer-survival-eda