In [64]:
```
# import the library
import numpy as np
import pandas as pd
import warnings
warnings.filterwarnings("ignore")
```

In [65]:
```
# Read the Csv
haberman = pd.read_csv("haberman.csv")
```

High Level Stastics

In [66]:
```
# Number of Points And Features
print(haberman.shape)   # 306 rows and 4 feature in our haberman dataset
```

```
(306, 4)
```

In [67]:
```
print(haberman.columns)
```

```
Index(['age', 'year', 'nodes', 'status'], dtype='object')
```

In [68]:
```
'''
independent Variable --- age,year,nodes
dependent Variable/target variable --- status
'''
```

Out[68]:
```
'\nindependent Variable --- age,year,nodes\ndependent Variable/target variabl
e --- status\n'
```

In [69]:
```
# number of class in target variable
#datapoint per class
print(haberman['status'].value_counts())
'''
Refrence Kaggle ---
1 = the patient survived 5 years or longer
2 = the patient died within 5 year
'''
# Two class we have in our Target Variable
```

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

Out[69]:
```
'\nRefrence Kaggle ---\n1 = the patient survived 5 years or longer\n2 = the p
atient died within 5 year\n'
```

In [70]:
```
#datapoint per class
'''
225 people surrvied after breast surgery more than five yaer means Suceesful
81 people died after surgery in five year means unsceesfull
225 people status is 1
81 people status is 2
looks like imbalanced dataset
'''
```

Out[70]:
```
'\n225 people surrvied after breast surgery more than five yaer means Suceesf
ul \n81 people died after surgery in five year means unsceesfull \n225 people
status is 1\n81 people status is 2\nlooks like imbalanced dataset\n'
```

In [71]:
```
print(haberman.head())
```

```
   age  year  nodes  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
```

DATA INFORMATION ---- we have 306 Patient Informtaion.225 People survived after breast surgery more than five yaer means Suceesful.And Rest of The 81 People died before 5 Year.And Final Point is Our Dataset is Not A Balanced DataSet.

In [ ]:

In [ ]:

Objective --- If a person has undergone the surgery,we have to tell that person is survived or not in 5 year based on some Features.

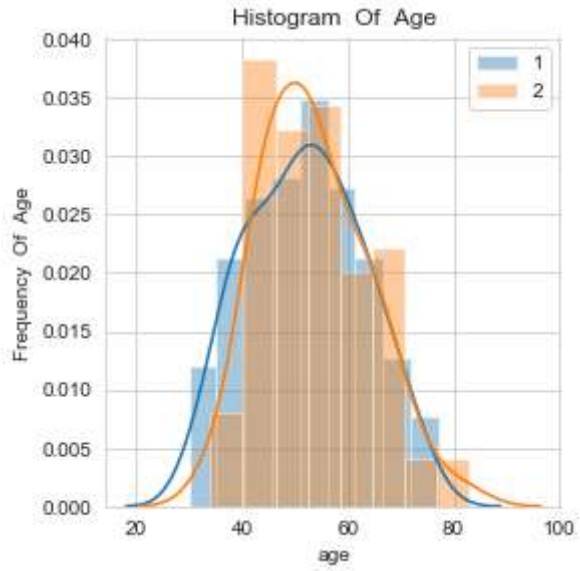
In [ ]:

In [ ]:

Univariate Analysis -- PDF, CDF, Boxplot, Voilin plots (Which Feature is More Important)

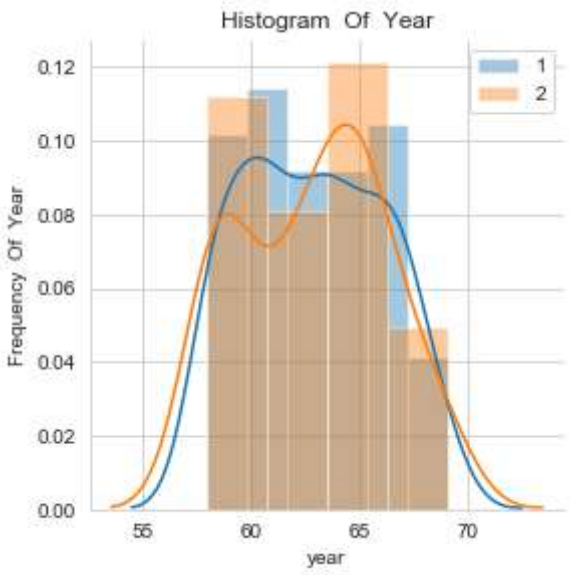
Histogram,CDF,PDF

In [106]:
```
# Histograms
import seaborn as sns
import matplotlib.pyplot as plt
sns.FacetGrid(haberman,hue='status',height=4).map(sns.distplot,'age')
plt.title('Histogram  Of  Age')
plt.ylabel('Frequency  Of  Age')
plt.legend()
plt.show()
```



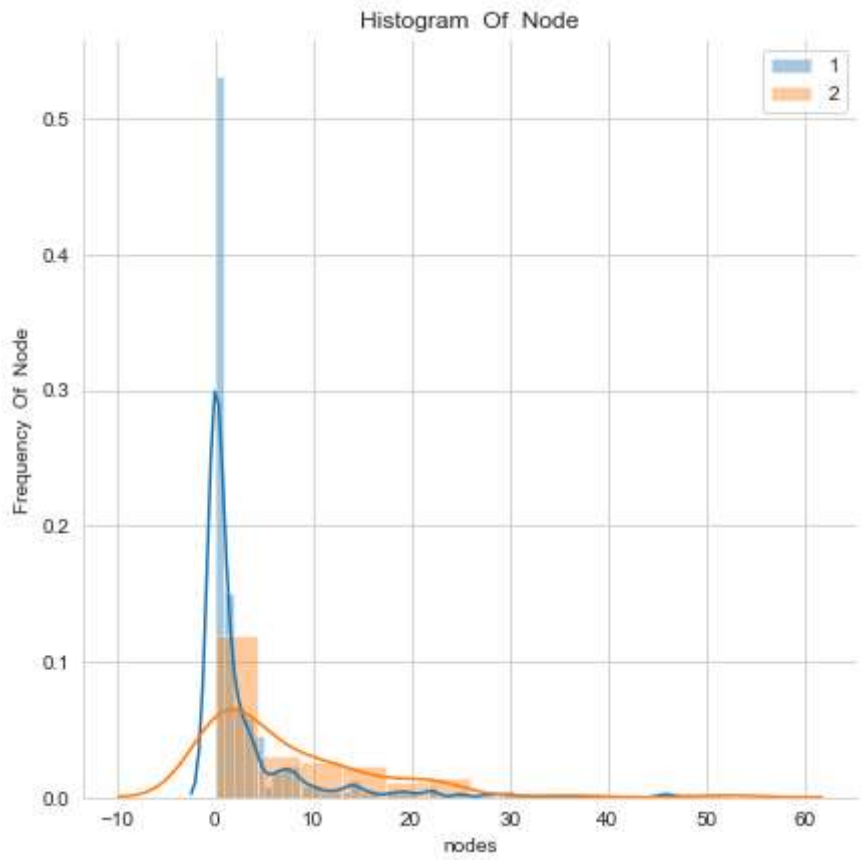
Conclusion -- Age is Not Important Feature.Age Frequency is Distribued Simmilarly Between Status-1 And Status-2.And Pdf is overlapping.

In [108]:
```
sns.FacetGrid(haberman,hue='status',height=4).map(sns.distplot,'year')
plt.title('Histogram  Of  Year')
plt.ylabel('Frequency  Of  Year')
plt.legend()
plt.show()
```



Conclusion -- Year is Not Important Feature.Year Frequency is Distribued Simmilarly Between Status-1 And Status-2.And Pdf is overlapping.

In [110]:
```
sns.FacetGrid(haberman,hue='status',height=6).map(sns.distplot,'nodes')
plt.title('Histogram  Of  Node')
plt.ylabel('Frequency  Of  Node')
plt.legend()
plt.show()
```

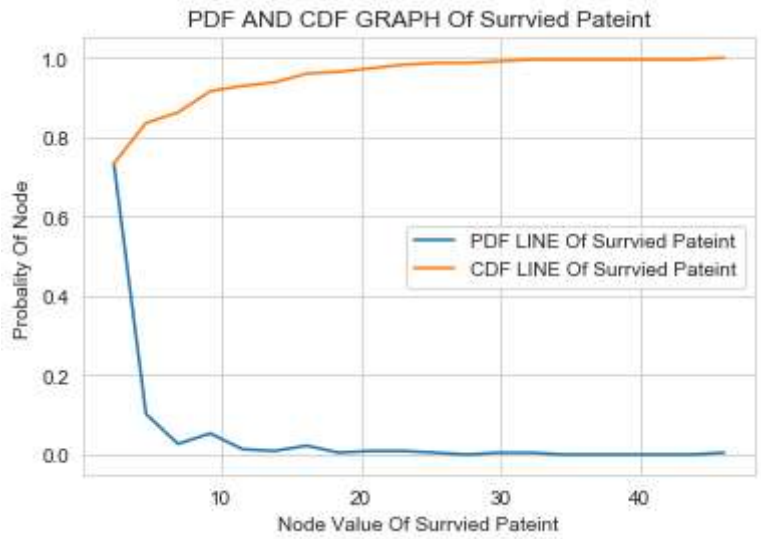


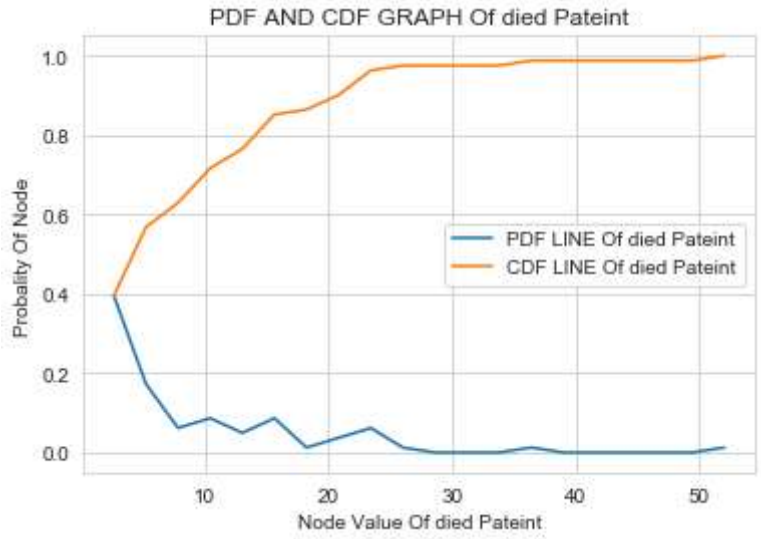
Conclusion -- Node is A Important Feature Here.If Node node values is less then higher chance of Patient Surrival But Node Value is high then less chance of Patient Surrival.

In [77]:
```
# i will Do more reserch on nodes Features
```

In [78]:
```
nodes_1 = haberman.loc[haberman['status']==1]
nodes_2 = haberman.loc[haberman['status']==2]
```

In [114]:
```
count,n_bins = np.histogram(nodes_1['nodes'],bins=20,density=True)
pdf = count/sum(count)
cdf = np.cumsum(pdf)
plt.plot(n_bins[1:],pdf,label='PDF LINE Of Survried Pateint')
plt.plot(n_bins[1:],cdf,label='CDF LINE Of Survried Pateint')
plt.title('PDF AND CDF GRAPH Of Survried Pateint')
plt.xlabel('Node Value Of Survried Pateint')
plt.ylabel('Probality Of Node')
plt.legend()
plt.show()
count,n_bins = np.histogram(nodes_2['nodes'],bins=20,density=True)
pdf = count/sum(count)
cdf = np.cumsum(pdf)
plt.plot(n_bins[1:],pdf,label='PDF LINE Of died Pateint')
plt.plot(n_bins[1:],cdf,label='CDF LINE Of died Pateint')
plt.title('PDF AND CDF GRAPH Of died Pateint')
plt.xlabel('Node Value Of died Pateint')
plt.ylabel('Probality Of Node')
plt.legend()
plt.show()
```





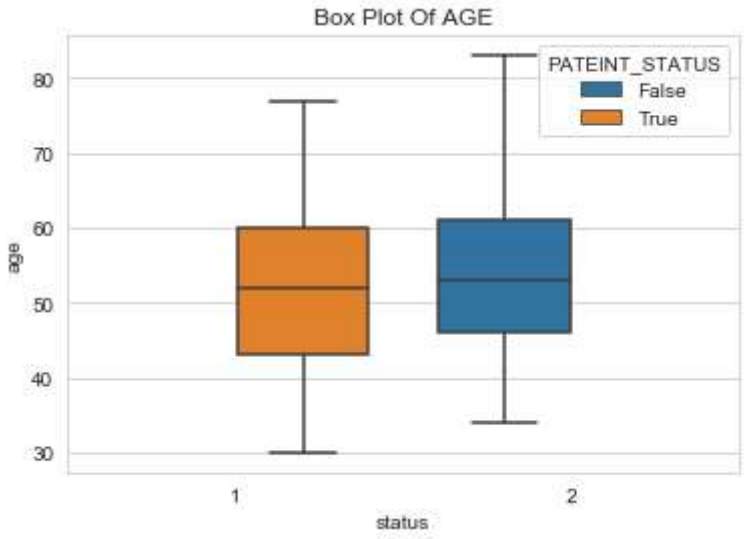
Conclusion --- If Node Value is Approximatly 5 then 70 Percent of Pateint Surrvied(Graph -1).Means Out of 225 Pateint 158 Pateint Surrvied.

If Node Value is Approximatly 5 then 40 Percent of Pateint Died (Graph -2).Means Out of 81 Pateint 36 Pateint Died.

BOXPLOT

In [125]:
```
import seaborn as sns
haberman["PATEINT_STATUS"] = haberman["status"].isin(["0", "1"])
bp1 = sns.boxplot(x = "status",y = "age",data=haberman,hue="PATEINT_STATUS")
plt.title('Box Plot Of AGE')
```

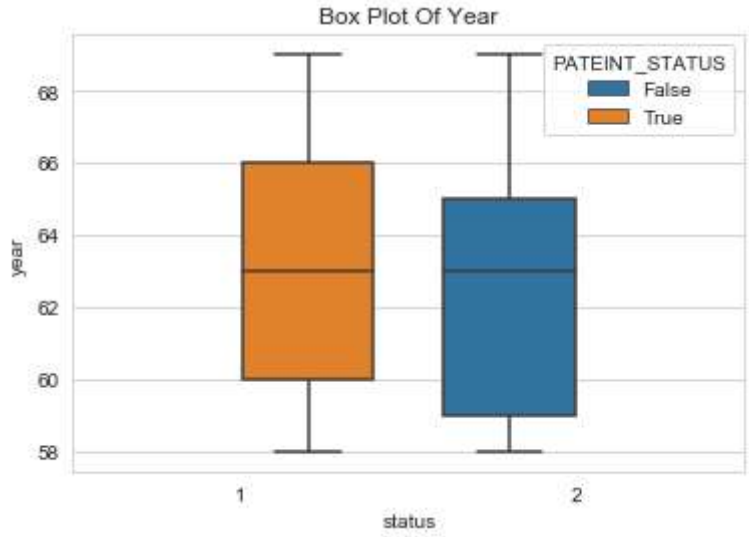
Out[125]: Text(0.5,1,'Box Plot Of AGE')



Conclusion --- Age is Not Important Feature.Age Distribution is Simmilar Between Target Variable.So i can't see Any Clear pattern

In [127]:
```
haberman["PATEINT_STATUS"] = haberman["status"].isin(["0", "1"])
sns.boxplot(x = "status",y = "year",data=haberman,hue="PATEINT_STATUS")
plt.title('Box Plot Of Year')
```

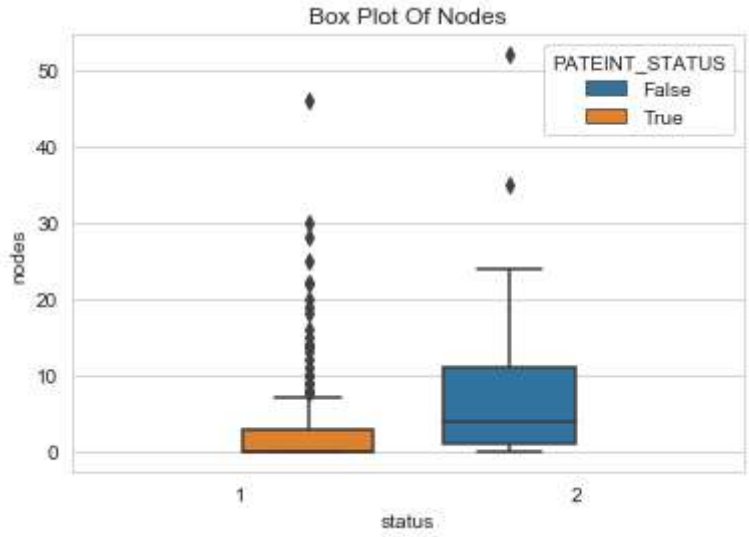
Out[127]: Text(0.5,1,'Box Plot Of Year')



Conclusion --- Yaer is Not Important Feature.if a Person Opertated after 65 then More Chance of Surrival.But i Can't See Any Clear Pattern.

In [128]:
```
haberman["PATEINT_STATUS"] = haberman["status"].isin(["0", "1"])
sns.boxplot(x = "status",y = "nodes",data=haberman,hue="PATEINT_STATUS")
plt.title('Box Plot Of Nodes')
```

Out[128]: Text(0.5,1,'Box Plot Of Nodes')

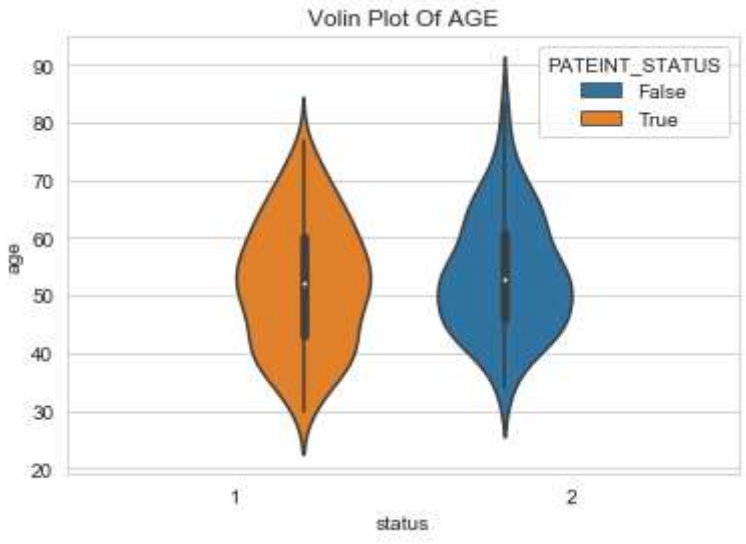


Conclusion --- Node Is A Important Feature.f Node Value is Less Then Higher Chance Of Surrival And If Node Value is High then Less Chance Of Surrvival.

Volin Plot

In [129]:
```
haberman["PATEINT_STATUS"] = haberman["status"].isin(["0", "1"])
sns.violinplot(x = "status",y = "age",data=haberman,hue="PATEINT_STATUS")
plt.title('Volin Plot Of AGE')
```

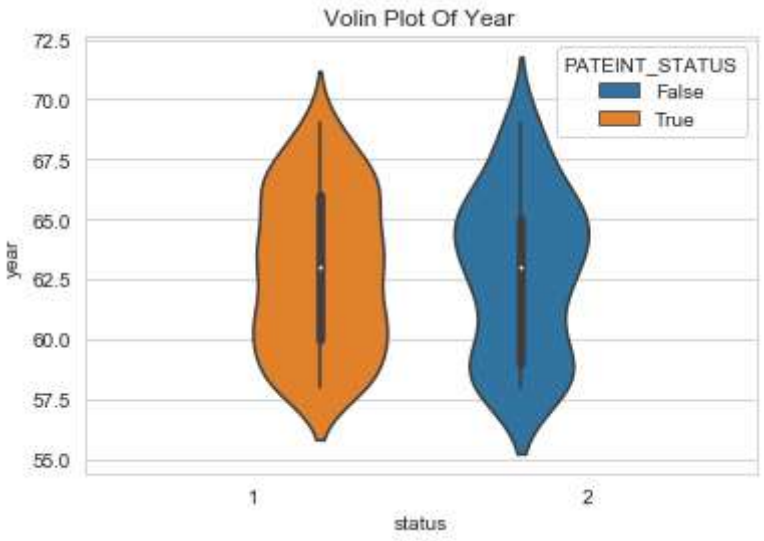
Out[129]: Text(0.5,1,'Volin Plot Of AGE')



Conclusion --- Age is Not Important Feature.Age Distribution is Simmilar Between Target Variable.So i can't see Any Clear pattern

In [131]:
```
haberman["PATEINT_STATUS"] = haberman["status"].isin(["0", "1"])
sns.violinplot(x = "status",y = "year",data=haberman,hue="PATEINT_STATUS")
plt.title('Volin Plot Of Year')
```

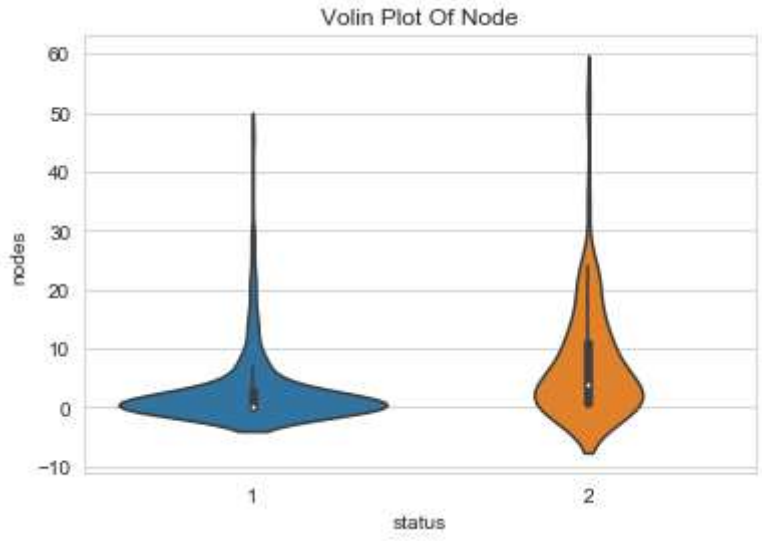
Out[131]: Text(0.5,1,'Volin Plot Of Year')



Conclusion --- Yaer is Not Important Feature.if a Person Opertated after 65 then More Chance of Surrival.But i Can't See Any Clear Pattern.

In [132]:
```
haberman["PATEINT_STATUS"] = haberman["status"].isin(["0", "1"])
sns.violinplot(x = "status",y = "nodes",data=haberman)
plt.title('Volin Plot Of Node')
```

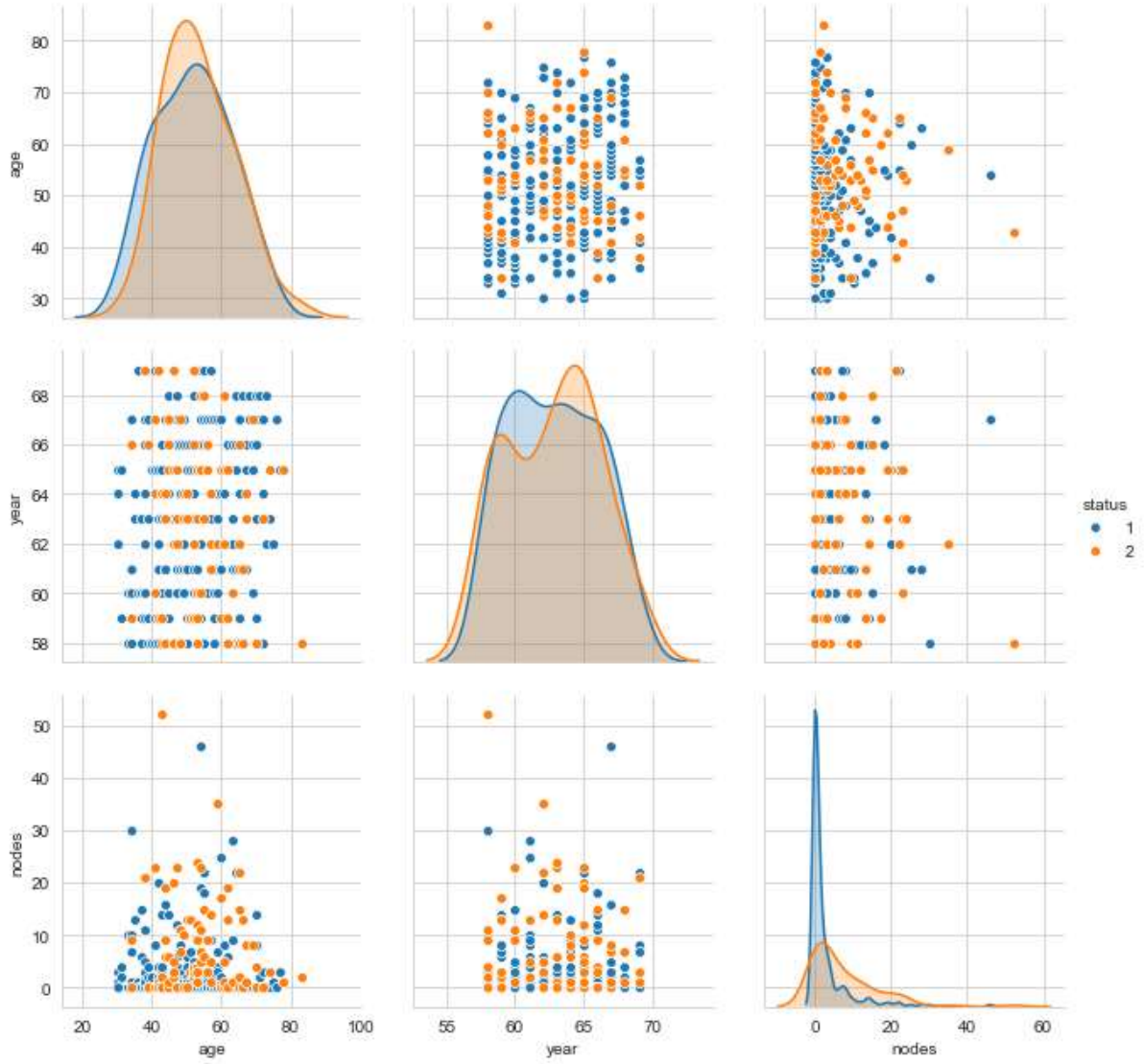
Out[132]: Text(0.5,1,'Volin Plot Of Node')



Conclusion --- Node Is A Important Feature.f Node Value is Less Then Higher Chance Of Surrival And If Node Value is High then Less Chance Of Surrvival.

BIVariate Analysis --- ScatterPlot And PairPlot

In [136]:
```
# Refrence --- https://seaborn.pydata.org/generated/seaborn.pairplot.html
sns.set_style("whitegrid")
sns.pairplot(haberman,hue="status",size=3,vars=['age', 'year', 'nodes'])
plt.show()
```



Conclusion --- I Can't See any Relationship Between Combination of two Variables.

In [ ]:
```

```

More Analysis on Node Value

In [137]:
```
print(np.percentile(nodes_1['nodes'],np.arange(0,100,10)))
'''
70 Percent Out Of 225 People if node value is less then 1---> 158 People
'''
```

```
[0. 0. 0. 0. 0. 0. 1. 2. 4. 8.]
```

Out[137]:
```
'\n70 Percent Out Of 225 People if node value is less then 1---> 158 People
\n'
```

In [138]:
```
print(np.percentile(nodes_2['nodes'],np.arange(0,100,10)))
'''
40 Percent Out Of 81 People if node value is less then 1---> 32 People
'''
```

```
[ 0.  0.  0.  1.  3.  4.  6.  9. 13. 20.]
```

Out[138]:
```
'\n40 Percent Out Of 81 People if node value is less then 1---> 32 People\n'
```

In [ ]:

In [ ]:

Final_Conclusion ---

1. We have 306 Pateint Information here.

2. Given Dataset is Unbalaced Dataset.

3. Most Important Feature is Node Value.

4. If Node value is less then 1 then 70% pateint Surrvied Means 158 out of 225.If i consider only Surrival pateint Dataset.

5. If Node value is less then 1 then 40% pateint Died Means 32 out of 81.If i consider only Died pateint Dataset.

6. So if i will randomly select a pateint and consider only node value then (306-32)*100/306 = 89% Chances is there i am Correct.

In [ ]:
```
S
```