In [1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

In [3]:
```
x=pd.read_csv("D:\\eda\\haberman new file.csv")
```

In [4]:
```
x.head()
```

Out[4]:

	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

In [5]:
```
print(x.shape)
```

```
(306, 4)
```

In [6]:
```
print(x.columns)
```

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

In [10]:
```
x["status"].value_counts()
```

Out[10]:
```
1    225
2     81
Name: status, dtype: int64
```

1.from this we can say that out of 306 patients 225 survived more than 5 years.And 81 patients survived less then 5 years.

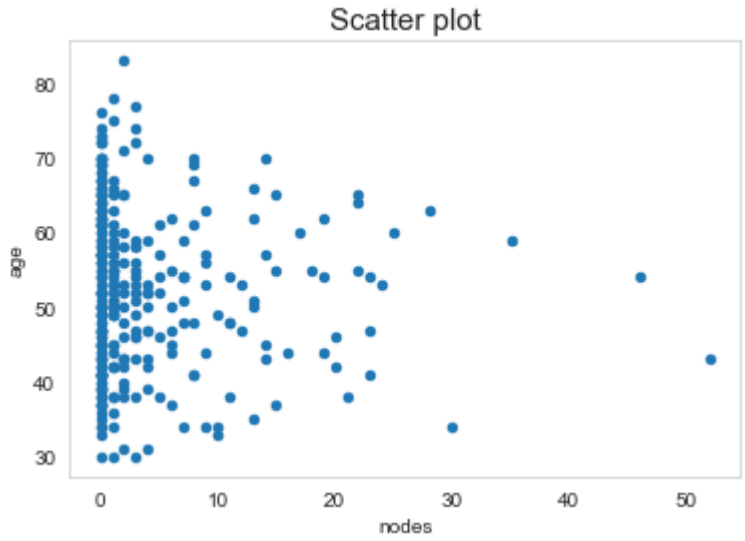
In [7]:
```
print(x.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306 entries, 0 to 305
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   age     306 non-null    int64
 1   year    306 non-null    int64
 2   nodes   306 non-null    int64
 3   status  306 non-null    int64
dtypes: int64(4)
memory usage: 9.7 KB
None
```

observations: 1.There are no missing values in our data set. 2.All the columns are of integer data type.

2-D Scatter plot:

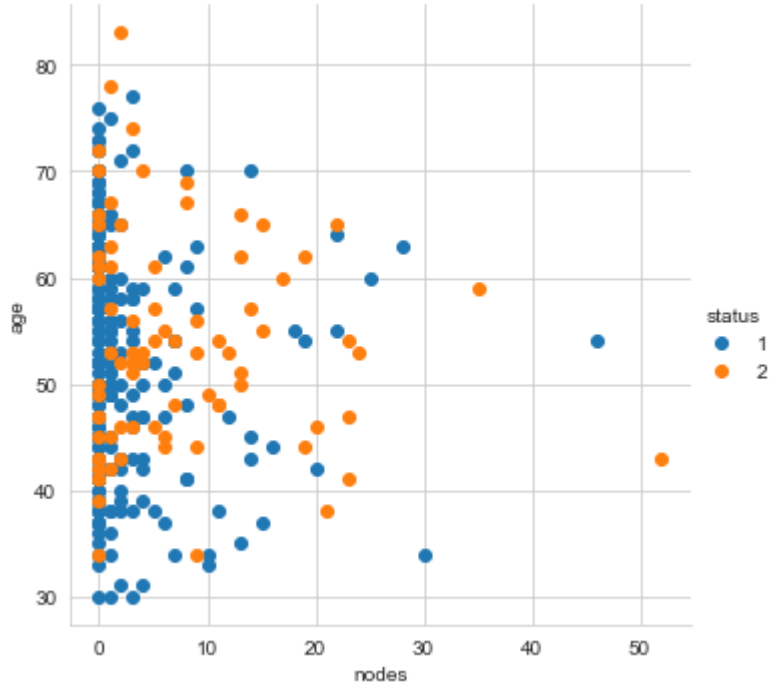
In [8]:
```
#now lets see with the help of plots.
x.plot(kind='scatter',x='nodes',y='age')
plt.title("Scatter plot",size=15)
plt.grid()
plt.show()
```



1.we can see that this plot overlapped and we can miss many data points,so this will not give us proper conclusion.

2-D Scatter plot with color-coding:

In [13]:
```
import warnings
warnings.filterwarnings("ignore")
sns.set_style('whitegrid')
sns.FacetGrid(x,hue='status',size=5)\
.map(plt.scatter,'nodes','age')\
.add_legend();
plt.show();
```



Observations 1.from this scatter plot we can able to distinguish between status 1 and 2.

2.blue dot shows survival more than 5 years.

3.orange dot shows survival less than 5 years.

4.this plot also cant give accurate conclusion as many points are overlapping.

5.Here status=1 mean patient survived more than 5 years.

6.And status=2 mean patient died within 5 year.

Pair-plot:

In [9]:
```
import warnings
warnings.filterwarnings("ignore")
plt.close();
sns.set_style("whitegrid");
sns.pairplot(x,hue='status',palette='flag',size=3,vars=['age','year','nodes']);
plt.show()
```

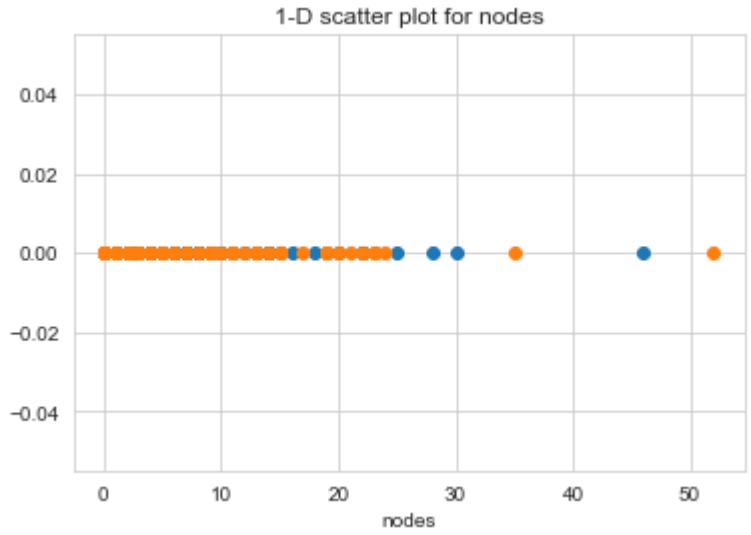
Observations: 1.we got 9 plots for data analyzation.

2.out of 9 plot we have to choose which plot can give us better conclusion

3.AGE and NODES are more usefull features this plot can provide conclusion more precisely.

4.I will take plot 3 age and nodes for further operations.

In [14]:
```
import numpy as np
x_long_survive=x.loc[x['status']==1]
x_short_survive=x.loc[x['status']==2]
plt.plot(x_long_survive["nodes"],np.zeros_like(x_long_survive['nodes']), 'o')
plt.plot(x_short_survive["nodes"],np.zeros_like(x_short_survive['nodes']),'o')
plt.title('1-D scatter plot for nodes')
plt.xlabel('nodes')
plt.show()
```



observations: 1.here we can say that the data of long survival and short survival gets overlapped so we can not get correct conclusions.

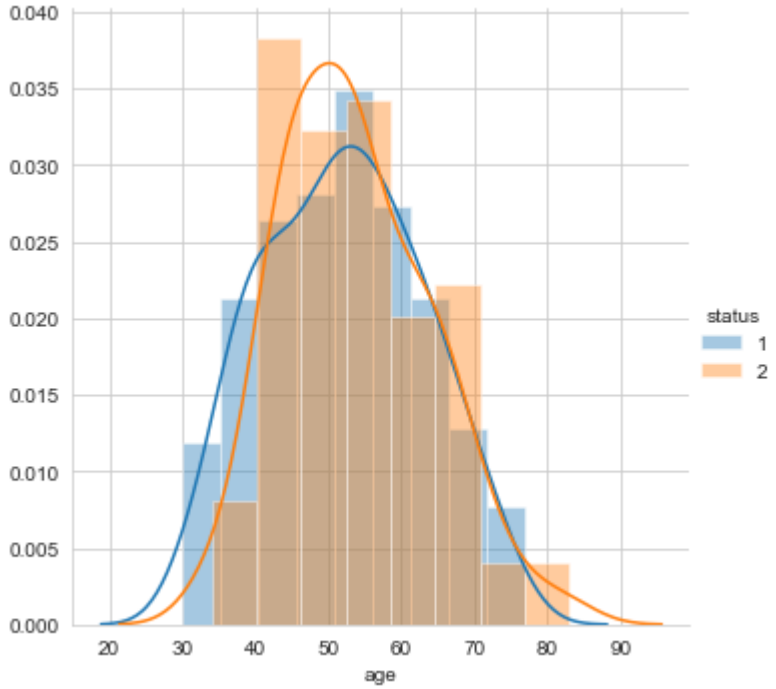
2.we can say very hard to make sense as points are overlapping.

Histogram,PDF(Probability density functions):

PDF of Age:

In [15]:
```
import warnings
warnings.filterwarnings("ignore")
sns.FacetGrid(x,hue="status", size=5)\
.map(sns.distplot,"age")\
.add_legend()
```

Out[15]: <seaborn.axisgrid.FacetGrid at 0x23473b11eb0>

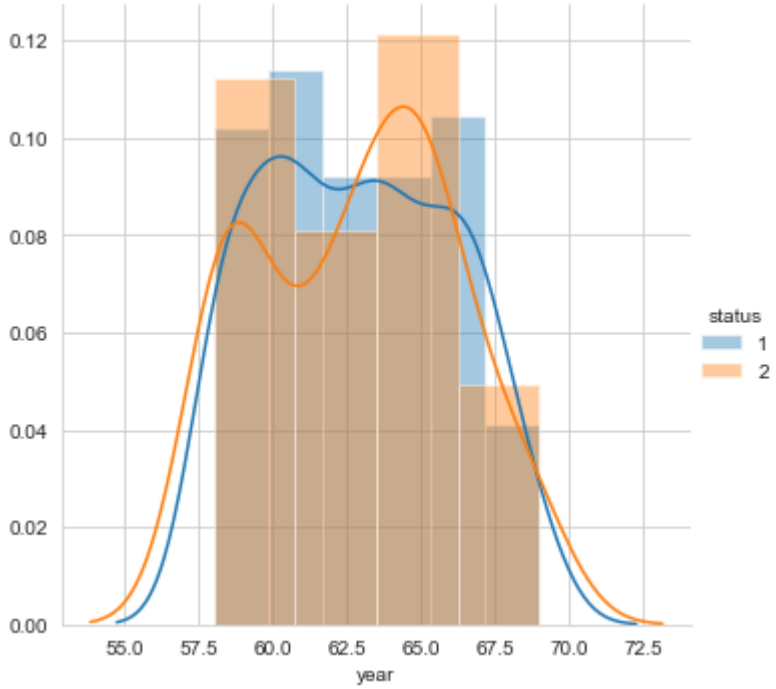


Observations: 1.Here i can see that people of age between 35 and 75 have the same survival and death. so its not possible to predict anything.

PDF of operation year:

In [16]:
```
import warnings
warnings.filterwarnings("ignore")
sns.FacetGrid(x,hue="status", size=5)\
.map(sns.distplot,"year")\
.add_legend()
```

Out[16]: <seaborn.axisgrid.FacetGrid at 0x23473b8e4c0>

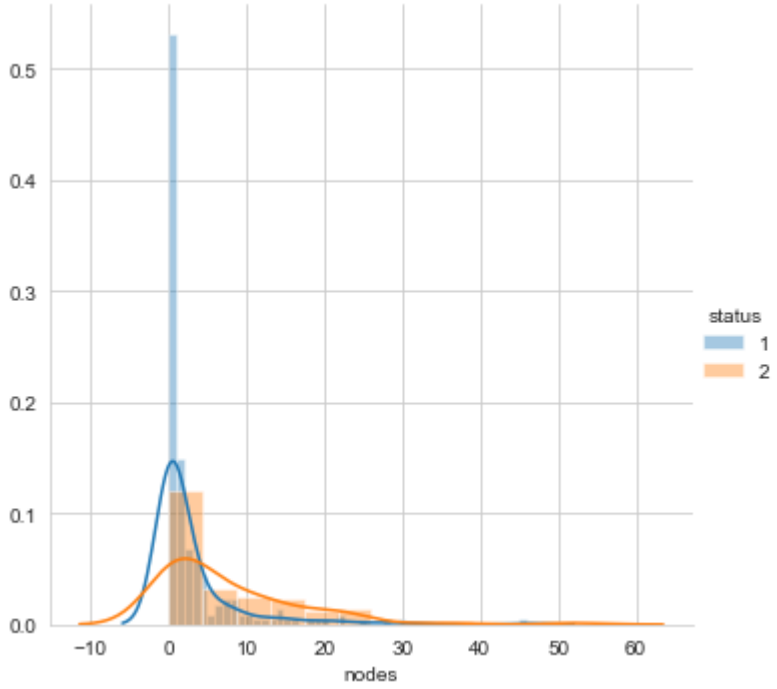


observations: 1.here also we can not predict anything as both pdfs overlap each other.

PDF of Nodes:

In [17]:
```
import warnings
warnings.filterwarnings("ignore")
sns.FacetGrid(x,hue="status", size=5)\
.map(sns.distplot,"nodes")\
.add_legend()
```

Out[17]: <seaborn.axisgrid.FacetGrid at 0x23473837a30>



Observations: 1.I can say that people with less nodes survive more.

2.people with more nodes survive less.

3.if patient have (nodes less than or equal to 0) patient will survive longest.

4.if patient have (nodes greater than 0 and nodes less than 4.) patient will survive longer

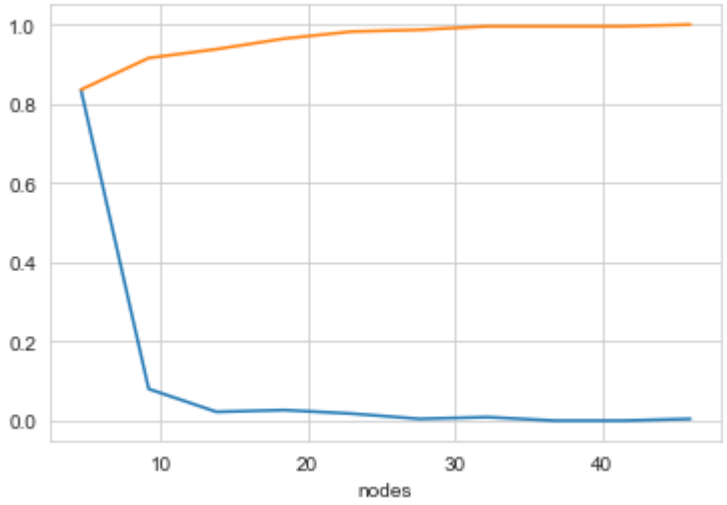
5.if patient have (nodes greater than 4) the survival of patient will be short. and chances of death is more.

CDF.(Cumulative distribution function):

CDF for long survival status:

In [18]:
```
counts, bin_edges = np.histogram(x_long_survive["nodes"], bins=10,
 density = True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges);
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf);
plt.plot(bin_edges[1:], cdf)
plt.xlabel('nodes')
plt.show()
```

```
[0.83555556 0.08       0.02222222 0.02666667 0.01777778 0.00444444
 0.00888889 0.         0.         0.00444444]
[ 0.   4.6  9.2 13.8 18.4 23.  27.6 32.2 36.8 41.4 46. ]
```



observations: 1.Here the CDF of long survival status on plot is shown by ORANGE color.

2.as we can see the orange line shows 82% chances of long survival if nodes detected is less than 5.

3.as we can clearly see as number of nodes increases chances of survival decreases.

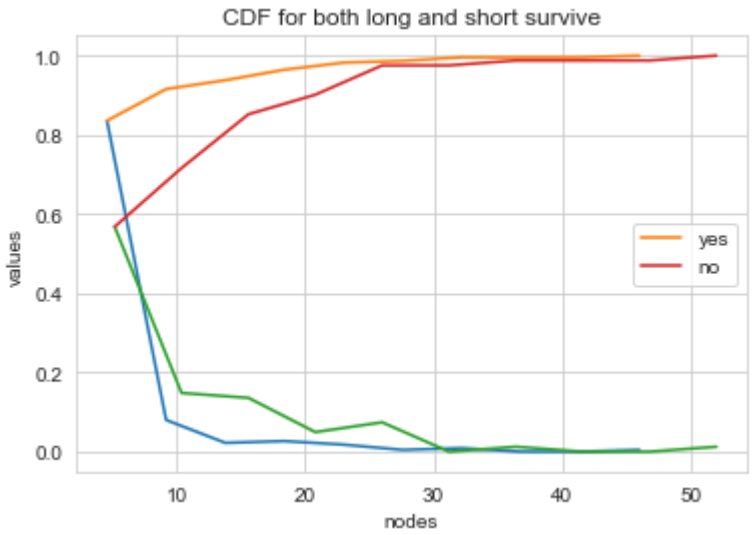
CDF for both long and short survival status:

In [21]:
```
counts, bin_edges = np.histogram(x_long_survive["nodes"], bins=10,
 density = True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges);
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf);
plt.plot(bin_edges[1:], cdf,label='yes')

counts, bin_edges = np.histogram(x_short_survive["nodes"], bins=10,
 density = True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges);


cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf);
plt.plot(bin_edges[1:], cdf,label='no')
plt.xlabel('nodes')
plt.ylabel('values')
plt.title('CDF for both long and short survive')
plt.legend()
plt.show()
```

```
[0.83555556 0.08       0.02222222 0.02666667 0.01777778 0.00444444
 0.00888889 0.         0.         0.00444444]
[ 0.   4.6  9.2 13.8 18.4 23.  27.6 32.2 36.8 41.4 46. ]
[0.56790123 0.14814815 0.13580247 0.04938272 0.07407407 0.
 0.01234568 0.         0.         0.01234568]
[ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
```



NOTE: 1.CDF for short survival is shown by RED line:

we can say from above cdf. 58 % people have nodes less then 5.

Mean,Standard Deviation:

In [46]:
```
print("Means:")
print(np.mean(x_long_survive["nodes"]))
print(np.mean(x_short_survive["nodes"]))


print("\nStandard Deviation:")
print(np.std(x_long_survive["nodes"]))
print(np.std(x_short_survive["nodes"]))




print("mean with outlier")
print(np.mean(np.append(x_long_survive["nodes"],50)))
print(np.mean(np.append(x_short_survive["nodes"],50)))
```

```
Means:
2.7911111111111113
7.45679012345679

Standard Deviation:
5.857258449412131
9.128776076761632
mean with outlier
3.0
7.975609756097561
```

.OBSERVATIONS: 1.we can see the mean of long survive is 2.79 and with outlier its 3 so there is no much diffrence.

2.the mean of short survive is 7.45 and with outlier its 7.97 there is more diffrence so we can say that the probability of short survive is more in dataset.

3.if we see the standard deviation for long survive its only 5.8 and for short survive its 9.2 so we can say the spread of data for short survive is more.

Median,Quantiles And percentiles:

In [48]:
```
print("Medians:")
print(np.median(x_long_survive["nodes"]))
print(np.median(np.append(x_long_survive["nodes"],50)));
print(np.median(x_short_survive["nodes"]))


print("\nQuantiles:")
print(np.percentile(x_long_survive["nodes"],np.arange(0, 100, 25)))
print(np.percentile(x_short_survive["nodes"],np.arange(0, 100, 25)))


print("\n90th Percentiles:")
print(np.percentile(x_long_survive["nodes"],90))
print(np.percentile(x_short_survive["nodes"],90))

from statsmodels import robust
print ("\nMedian Absolute Deviation")
print(robust.mad(x_long_survive["nodes"]))
print(robust.mad(x_short_survive["nodes"]))
```

```
Medians:
0.0
0.0
4.0

Quantiles:
[0. 0. 0. 3.]
[ 0.  1.  4. 11.]

90th Percentiles:
8.0
20.0

Median Absolute Deviation
0.0
5.930408874022408
```

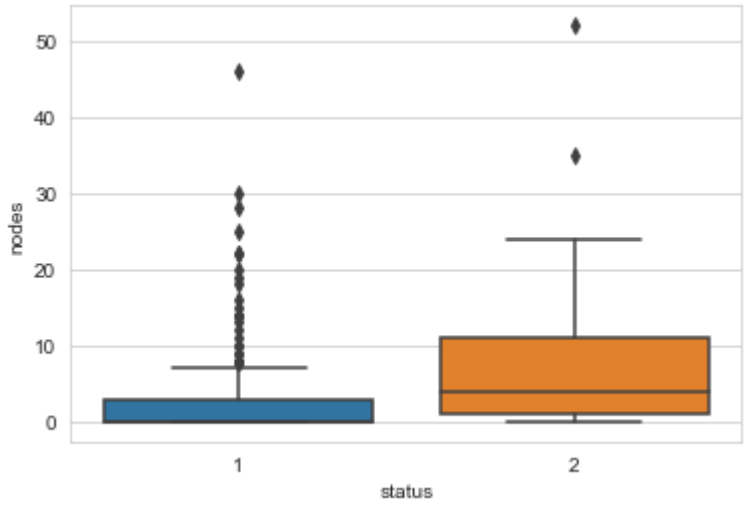
observations: 1.from above observation its clear that averege nodes in long survival is 0. And for short survival it is 4.

2.our quantiles shows that in long survival 50% nodes are 0.

3.75% of patients have less than 3 nodes.

Box plot and Whiskers:

In [8]:
```
sns.boxplot(x="status",y="nodes",data=x)
plt.show()
```

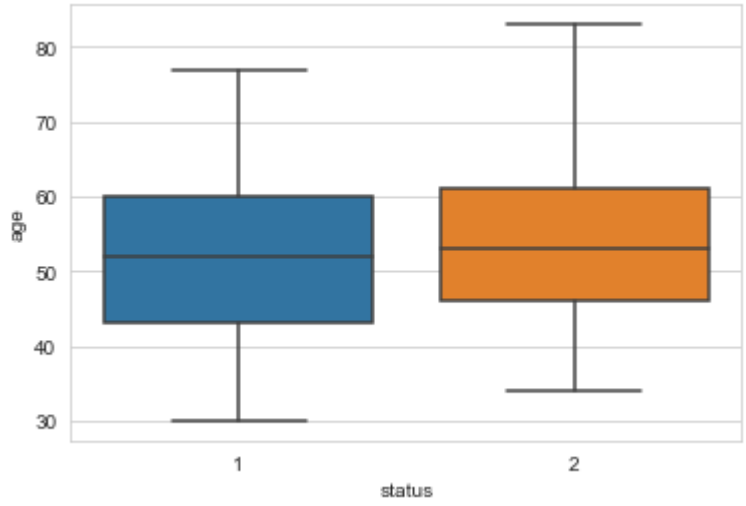


observations:

1.box plot clearly shows that if patients have less number of nodes they survived more and vice versa.

2.patients who have nodes more than 20 died.

In [9]:
```
sns.boxplot(x="status",y="age",data=x)
plt.show()
```



observation:

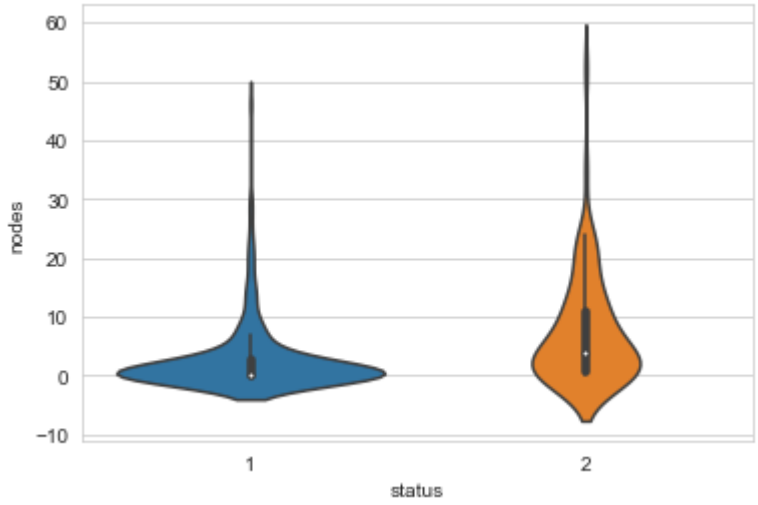
1.patients of age between 40-60 likely survive for long period.

2.patients between age 48-64 likely to survive for short period.

3.we can say as age increases chances of survival decreases.

Violin plots:

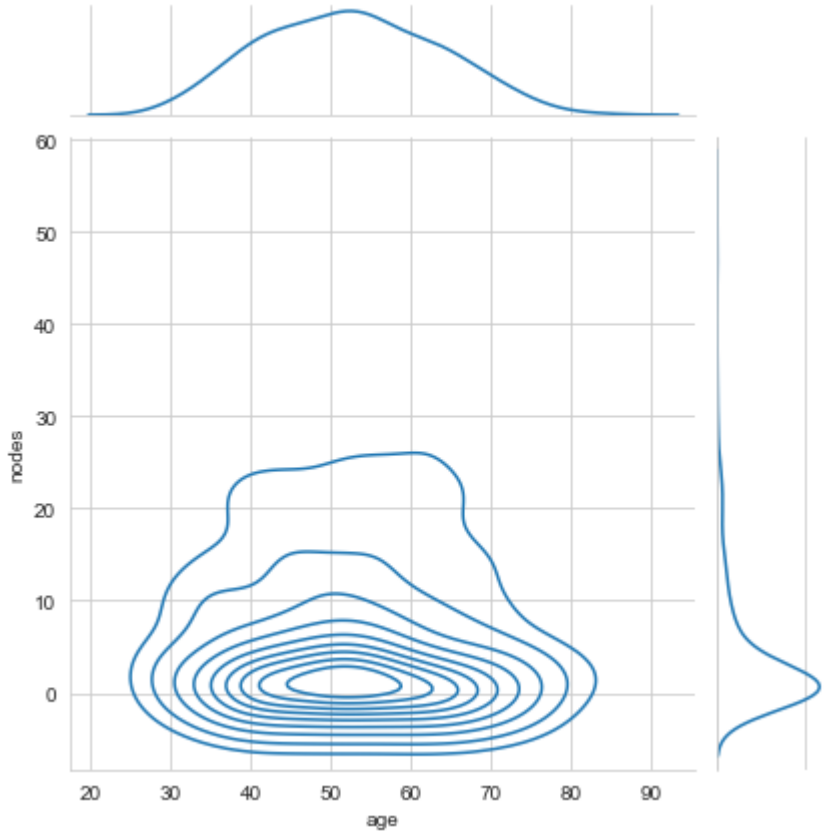
In [13]:
```
sns.violinplot(x='status',y='nodes',data=x,size=9)
plt.show()
```



observations: 1.from violin plot we can clearly see that most of the patients who survived had nodes 0 or less than 0.

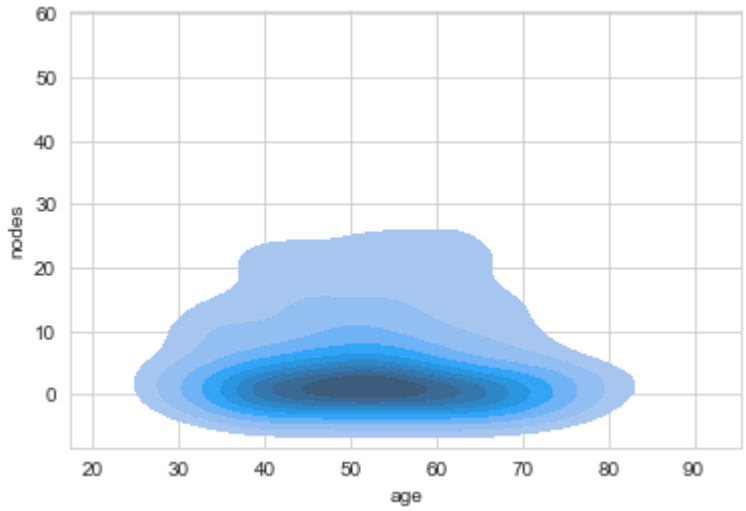
Multivariate probability density, contour plot:

In [17]:
```
sns.jointplot(x="age",y="nodes",data=x,kind="kde")
plt.grid()
plt.show()
```



In [21]:
```
sns.kdeplot(data=x , x='age',y='nodes',fill=True)
```

Out[21]:
```
<AxesSubplot:xlabel='age', ylabel='nodes'>
```



observation: here i can say that density of long survival is more from age range 45-55 and nodes 0-3

conclusion: 1.we can clearly say that if patients have 0 nodes they likely to survive longest.

2.patients with nodes greater than 10 are likely to dead.

3.from box plot we can say that as number of nodes increases chances of survival decreases.

4.nodes are the most important for analysation. most of patients likely to die if they have nodes greater than or equal to 2.