In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sbn
import matplotlib.pyplot as plt
hb=pd.read_csv('haberman.csv')
hb
```

Out[1]:

	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
...	...	...	...	...
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 × 4 columns

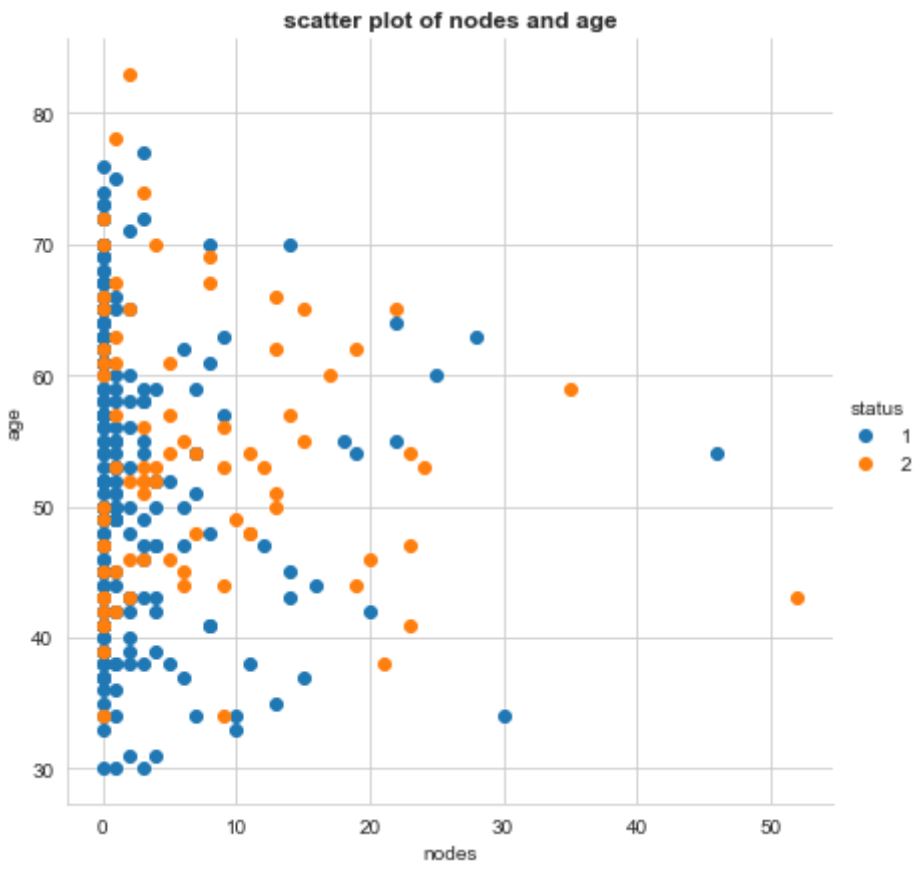
In [2]:
```
hb.status.value_counts()
```

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

Observation:

1. It's a imbalanced data set as the whole data set tells that 225 patients have survived 5 years and longer,whereas 81 patients has survived less than 5 years

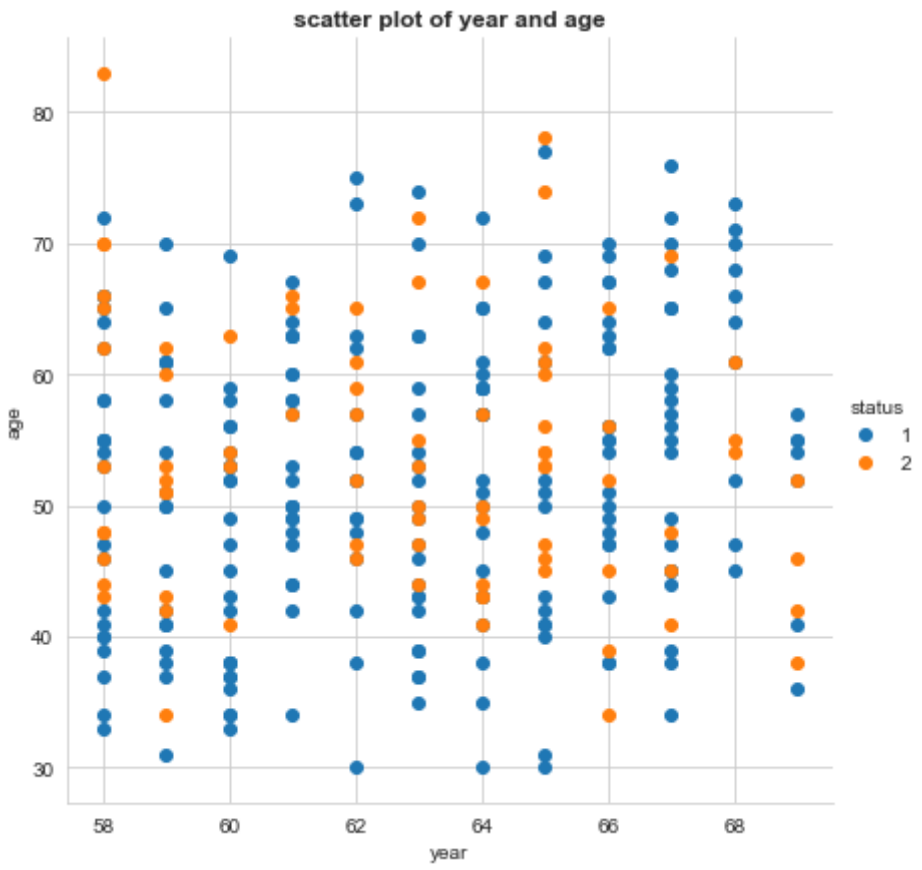
In [3]:
```
sbn.set_style("whitegrid") #Bivariate analysis
sbn.FacetGrid(hb,hue='status',height=6).map(plt.scatter,'nodes','age').add_legend()
plt.title('scatter plot of nodes and age',fontweight='bold')
plt.show()
```



observation:

1. There are more number of patients who have less than 10 nodes.
2. Status 1 patients are comparatively very high than status 2 pateints in patients of zero nodes.
3. Cant distinguish between patients of status 1 and status 2 as they have ovelaped.

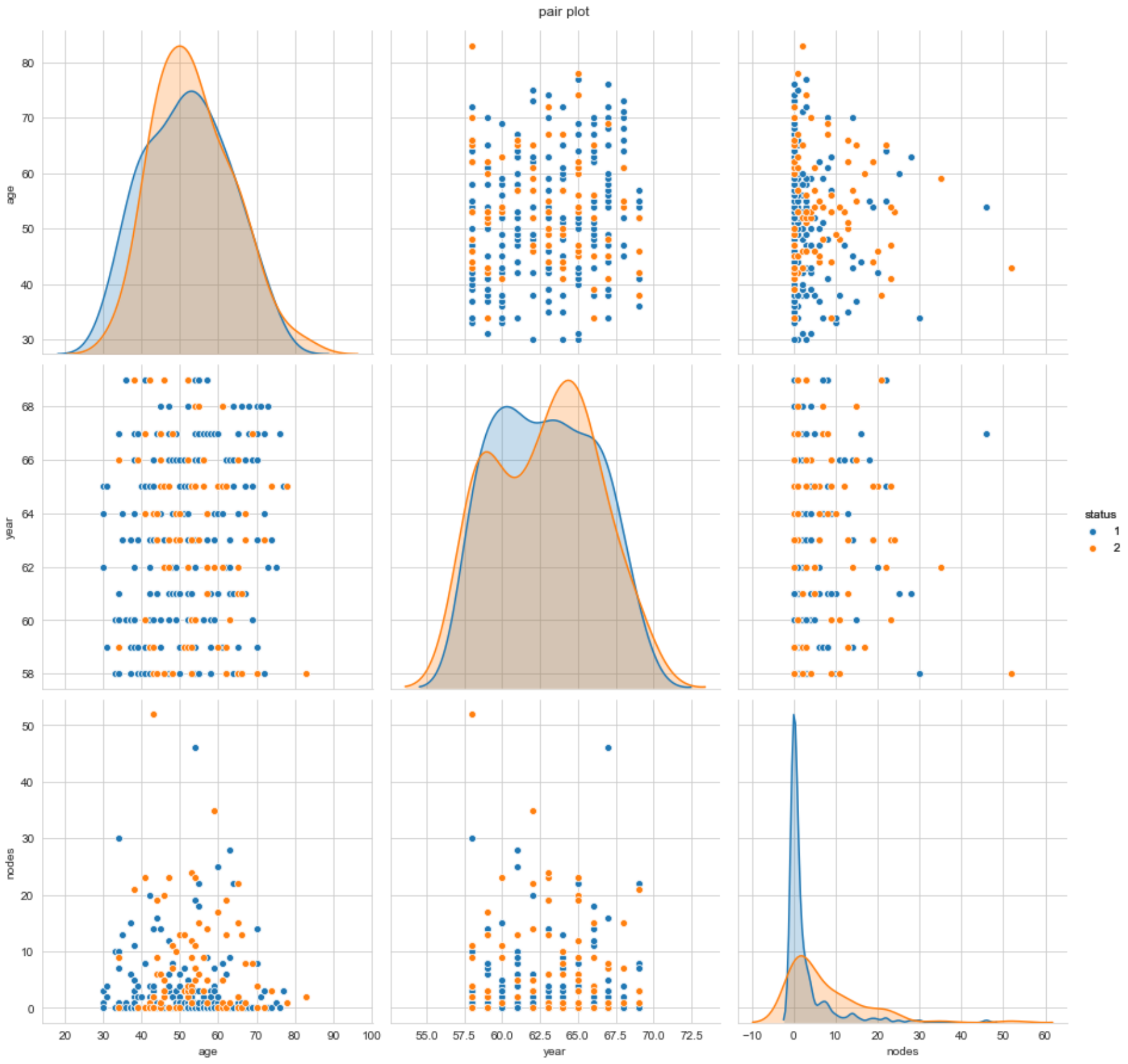
In [4]:
```
sbn.set_style("whitegrid") #Bivariate analysis
sbn.FacetGrid(hb,hue='status',height=6).map(plt.scatter,'year','age').add_legend()
plt.title('scatter plot of year and age',fontweight='bold')
plt.show()
```



observation:

1. the age group of 30-40 has more survival chances
2. while other age groups couldn't distinguish between survival and non survival patients

In [5]:
```
sbn.set_style("whitegrid") #Bivariate analysis
p=sbn.pairplot(hb,hue='status',height=4).add_legend()
p.fig.suptitle('pair plot',y=1.02)
plt.show()
```

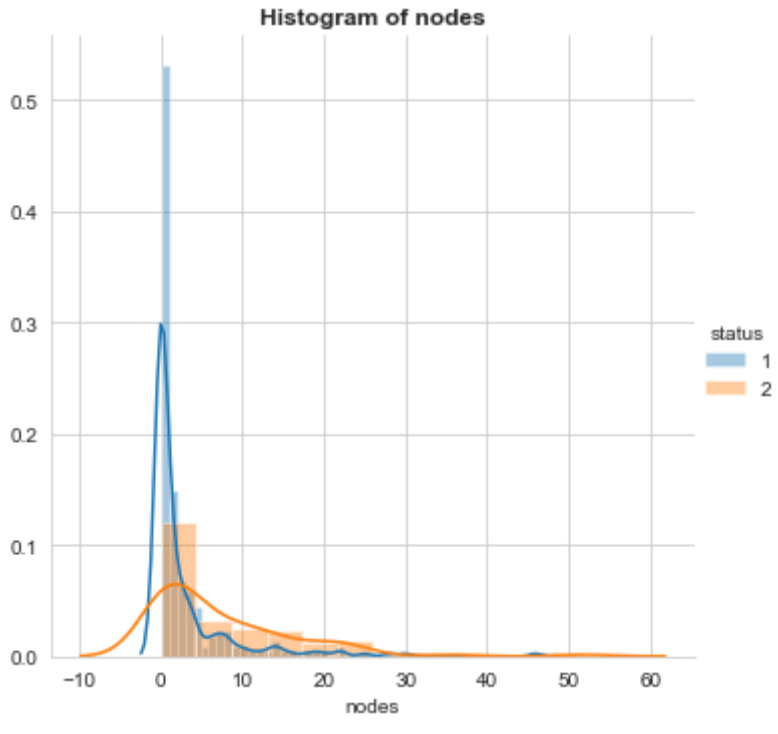


observation:

1. All the plot were totally overlaped
2. the plots of year Vs nodes seems better

In [6]:
```
sbn.FacetGrid(hb,hue='status',height=5).map(sbn.distplot,'nodes').add_legend() #univariate analysis
plt.title('Histogram of nodes',fontweight='bold')
```

Out[6]: Text(0.5, 1.0, 'Histogram of nodes')

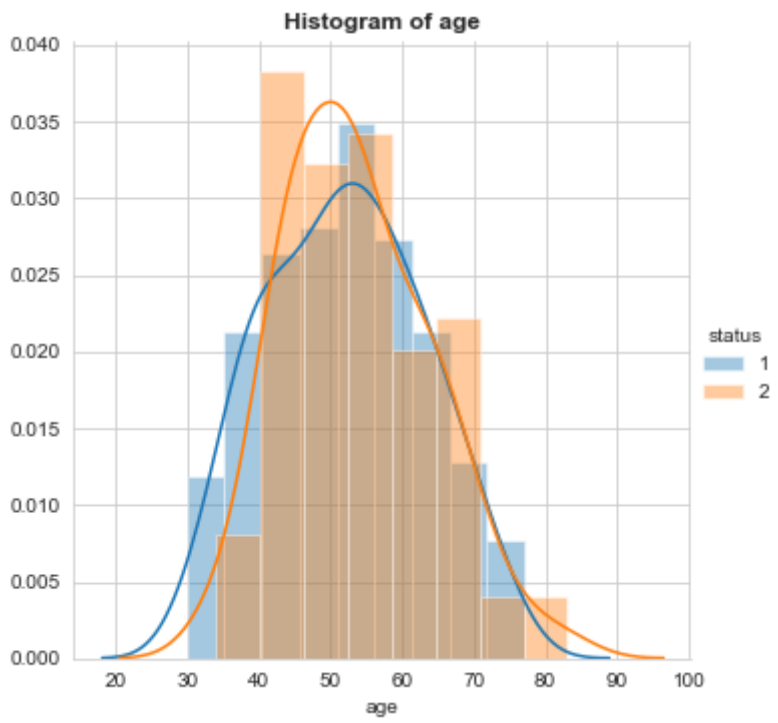


observation:

1. Both status 1 and status 2 plots are right skewed.
2. survival rate of patients with zero nodes are very high
3. patiemts with higher nodes have survived less

In [7]:
```
sbn.FacetGrid(hb, hue="status", height=5).map(sbn.distplot, "age").add_legend(); #univariate analysis
plt.title('Histogram of age',fontweight='bold')
```

Out[7]: Text(0.5, 1.0, 'Histogram of age')

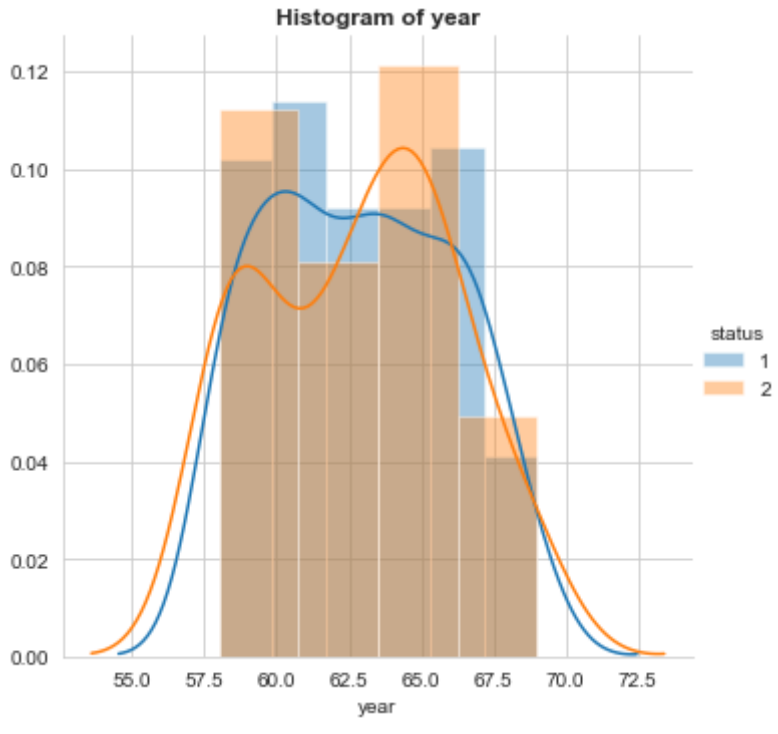


Observation:

1. both plots seems to be normally distrubuted and are overlaped.
2. the age group of 40-46 yrs has less survived

In [8]:
```
sbn.FacetGrid(hb, hue="status", height=5).map(sbn.distplot, "year").add_legend(); #univariate analysis
plt.title('Histogram of year',fontweight='bold')
```

Out[8]: Text(0.5, 1.0, 'Histogram of year')



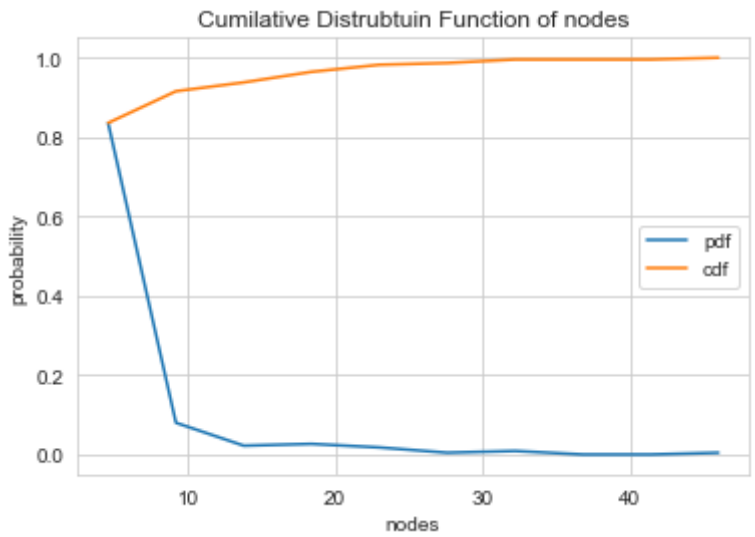
Observation:

1. In the year 1960,highest number of patients suvived
2. In the year 1965,lowest number of patients suvived
3. The plots of both categories are overlaped.

In [9]:
```
hb_status1=hb.loc[hb['status']==1]
hb_status2=hb.loc[hb['status']==2]
```

In [10]:
```
counts,binedges=np.histogram(hb_status1['nodes'],bins=10,density=True)
pdf=counts/sum(counts)
cdf=np.cumsum(pdf)
p,=plt.plot(binedges[1:],pdf)
c,=plt.plot(binedges[1:],cdf)
plt.xlabel('nodes')
plt.ylabel('probability')
plt.title('Cumilative Distrubtuin Function of nodes')
plt.legend([p,c],['pdf','cdf'])
```

Out[10]: <matplotlib.legend.Legend at 0x18f5aff6ca0>

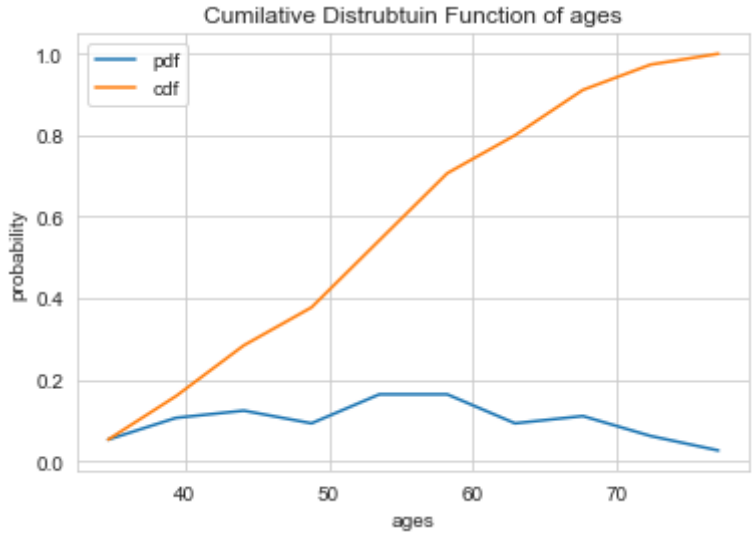


Observation:

1. 82% of survived patients are with zero nodes
2. Survived people are very few who have high number of nodes

In [11]:
```
counts,binedges=np.histogram(hb_status1['age'],bins=10,density=True)
pdf=counts/sum(counts)
cdf=np.cumsum(pdf)
p,=plt.plot(binedges[1:],pdf)
c,=plt.plot(binedges[1:],cdf)
plt.xlabel('ages')
plt.ylabel('probability')
plt.title('Cumilative Distrubtuin Function of ages')
plt.legend([p,c],['pdf','cdf'])
```

Out[11]: <matplotlib.legend.Legend at 0x18f5b69bfd0>



observation:

1. The 50-60 age group are comparatively more in number
2. the plot has been keep raising.

In [13]:
```
print('survived pateints mean of nodes:',np.mean(hb_status1['nodes']))
print('non survived pateints mean of nodes:',np.mean(hb_status2['nodes']))
print('survived pateints mean of age:',np.mean(hb_status1['age']))
print('non survived pateints mean of age:',np.mean(hb_status2['age']))
```

```
survived pateints mean of nodes: 2.7911111111111113
non survived pateints mean of nodes: 7.45679012345679
survived pateints mean of age: 52.01777777777778
non survived pateints mean of age: 53.67901234567901
```

observation:

1. There is much differnce in mean of nodes of survived patients Vs non survived patients.
2. Therfore we can tell that patients of less nodes are more likely to surivive.
3. There is no much differnce in mean of age of survived patients Vs non survived patients.
4. Therfore,we couldn't tell distinguish between survived and non survived patients

In [14]:
```
print('standard deviation of nodes of survived ptatients',np.std(hb_status1['nodes']))
print('standard deviation of nodes of non-survived ptatients',np.std(hb_status2['nodes']))
print('standard deviation of age of survived ptatients',np.std(hb_status1['age']))
print('standard deviation of age of non-survived ptatients',np.std(hb_status2['age']))
```

```
standard deviation of nodes of survived ptatients 5.857258449412131
standard deviation of nodes of non-survived ptatients 9.128776076761632
standard deviation of age of survived ptatients 10.98765547510051
standard deviation of age of non-survived ptatients 10.10418219303131
```

observation:

1. There is much differnce in standard deviation of nodes of survived patients Vs non survived patients.
2. Therfore we can tell that patients of less nodes are more likely to surivive.
3. There is no much differnce in standard deviation of age of survived patients Vs non survived patients.
4. Therfore,we couldn't tell distinguish between survived and non survived patients

In [15]:
```
print(np.median(hb_status1['nodes']))
print(np.median(hb_status2['nodes']))
print(np.median(hb_status1['age']))
print(np.median(hb_status2['age']))
```

```
0.0
4.0
52.0
53.0
```

Observation:

1. The median of nodes of survived patients is 0,whereas mean is 2.79 which tells,there are many outliers
2. The median of nodes of non-survived patients is 2,therefoe tells patiemts with nodes are less likely to survive
3. There is no much differnce in Median of age of survived patients Vs non survived patients.

In [16]:
```
print(np.percentile(hb_status1['nodes'],np.arange(25,125,25)))
print(np.percentile(hb_status2['nodes'],np.arange(25,125,25)))
```

```
[ 0.  0.  3. 46.]
[ 1.  4. 11. 52.]
```

Observation:

1. In both cases,The 100th percentile value is far greater than 75th and 50th percentiles,which tells there are outliers.

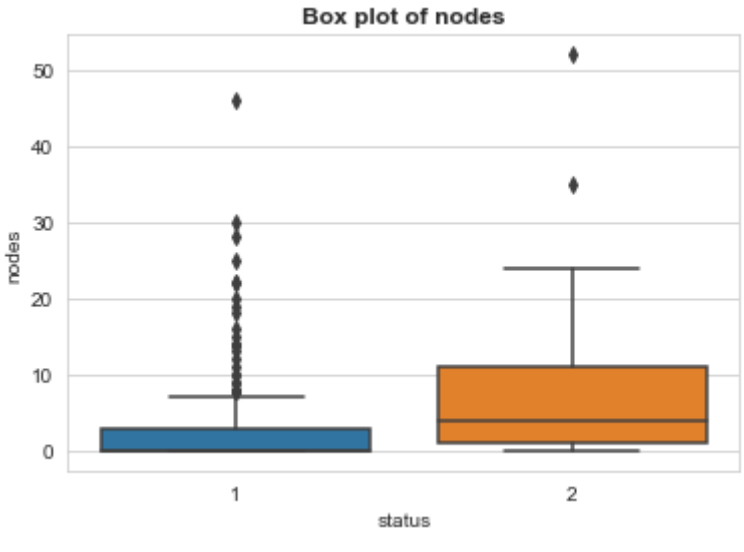
In [17]:
```
print(np.percentile(hb_status1['age'],np.arange(25,125,25)))
print(np.percentile(hb_status2['age'],np.arange(25,125,25)))
```

```
[43. 52. 60. 77.]
[46. 53. 61. 83.]
```

Observation:

1. There is no much differnce in percentiles of age of survived patients Vs non survived patients.

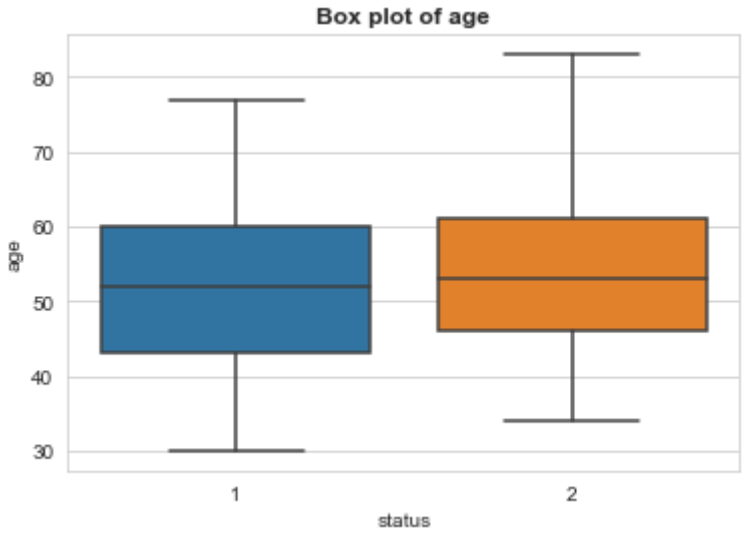
In [18]:
```
sbn.boxplot(x='status',y='nodes',data=hb)
plt.title('Box plot of nodes',fontweight='bold')
plt.show()
```



Observation:

1. patients who survived have nodes less than 8,whereas there are many outliers
2. 75% of patients who didnt survived has less than 12 nodes

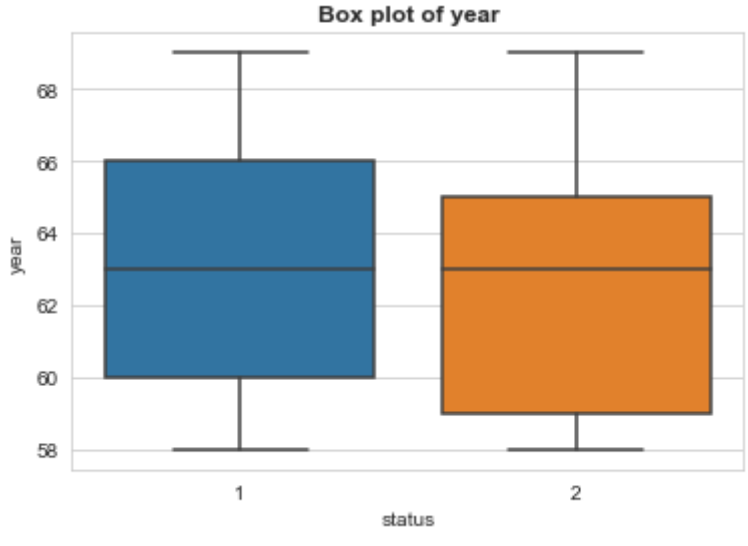
In [19]:
```
sbn.boxplot(x='status',y='age',data=hb)
plt.title('Box plot of age',fontweight='bold')
plt.show()
```



Observation:

1. In both survived and non survived categories, the 25th and 75th percentiles are of 43 - 62 age groups.

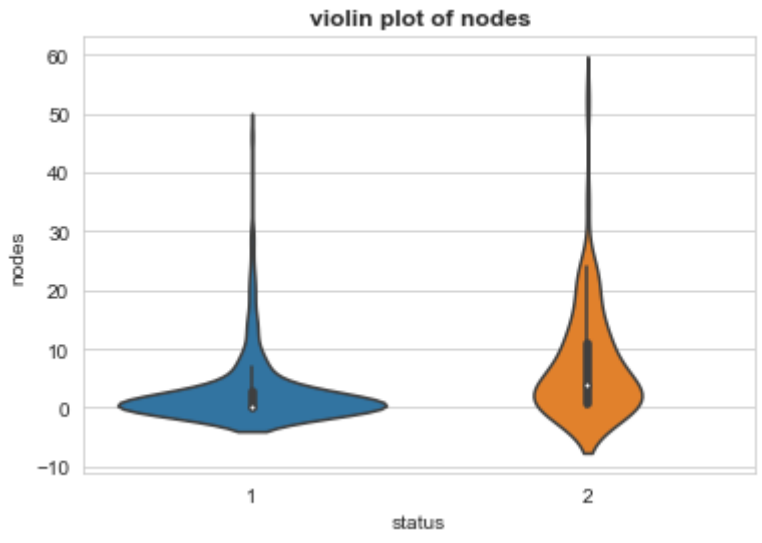
In [20]:
```
sbn.boxplot(x='status',y='year',data=hb)
plt.title('Box plot of year',fontweight='bold')
plt.show()
```



Observation:

1. There isnt much differnce between years of both survived and non survived categories.

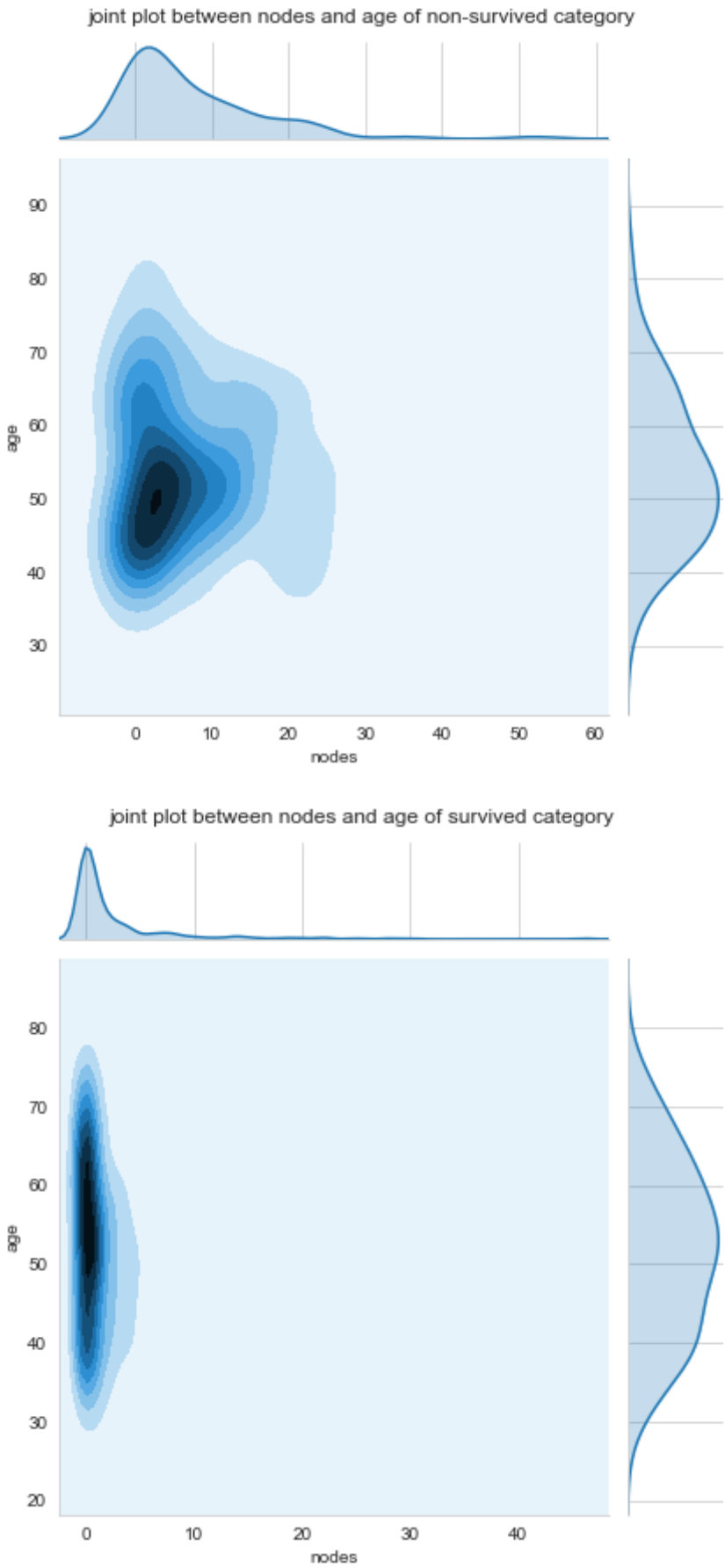
In [21]:
```
sbn.violinplot(x='status',y='nodes',data=hb,height=6)
plt.title('violin plot of nodes',fontweight='bold')
plt.show()
```



Observation:

1. There are many patients in with zero nodes of both survived and non survived categories.
2. Outliers are comparatively high in survived case.

In [22]:
```
s2=sbn.jointplot(x="nodes", y="age", data=hb_status2, kind="kde");
s2.fig.suptitle('joint plot between nodes and age of non-survived category',y=1.02)
plt.show();
s1=sbn.jointplot(x="nodes", y="age", data=hb_status1, kind="kde");
s1.fig.suptitle('joint plot between nodes and age of survived category',y=1.02)
plt.show();
```



Observation:

1. The age group of 40 to 60 with higher number of nodes(20) has less survived,whereas the age group of 40-60 with less number of nodes(3) has survived more.

Conclusion:

1. The age factor couldnt predict the survival nature,but the age group of 30-40 has survived more.
2. The year factor couldnt predict the survival nature.
3. lesser the number of nodes,more likely to survive.
4. couldnt classify between survival and non survival ppl, as it is a imbalanced data set.