Attribute Information:

Age of patient at time of operation (numerical)

Patient's year of operation (year - 1900, numerical)

Number of positive axillary nodes detected (numerical)

Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the patient died within 5 years

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

In [2]:
```
haber = pd.read_csv('haberman.csv')
```

In [3]:
```
haber.shape
```

Out[3]:
```
(306, 4)
```

In [4]:
```
haber.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306 entries, 0 to 305
Data columns (total 4 columns):
age       306 non-null int64
year      306 non-null int64
nodes     306 non-null int64
status    306 non-null int64
dtypes: int64(4)
memory usage: 9.6 KB
```

In [5]:
```
haber.describe()
```

Out[5]:

	age	year	nodes	status
count	306.000000	306.000000	306.000000	306.000000
mean	52.457516	62.852941	4.026144	1.264706
std	10.803452	3.249405	7.189654	0.441899
min	30.000000	58.000000	0.000000	1.000000
25%	44.000000	60.000000	0.000000	1.000000
50%	52.000000	63.000000	1.000000	1.000000
75%	60.750000	65.750000	4.000000	2.000000
max	83.000000	69.000000	52.000000	2.000000

In [6]:
```
haber.head()
```

Out[6]:

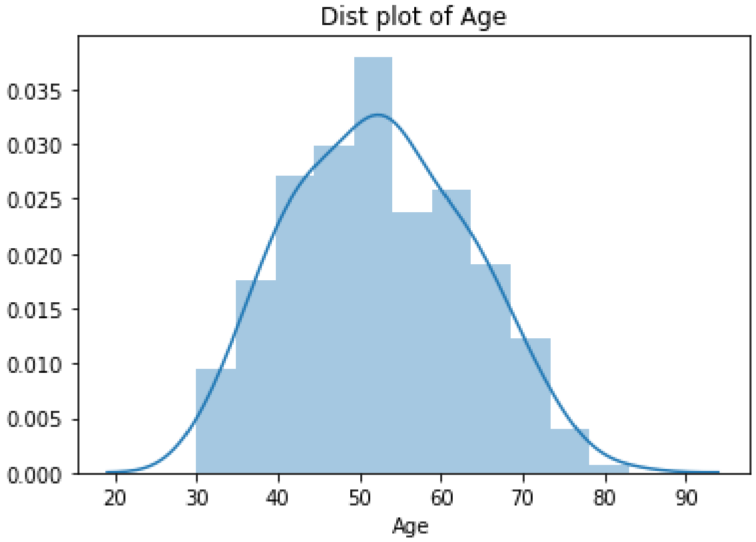
	age	year	nodes	status
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1

Univariate

Univariate_Analysis of Age

In [7]:
```
#Dist plot
```

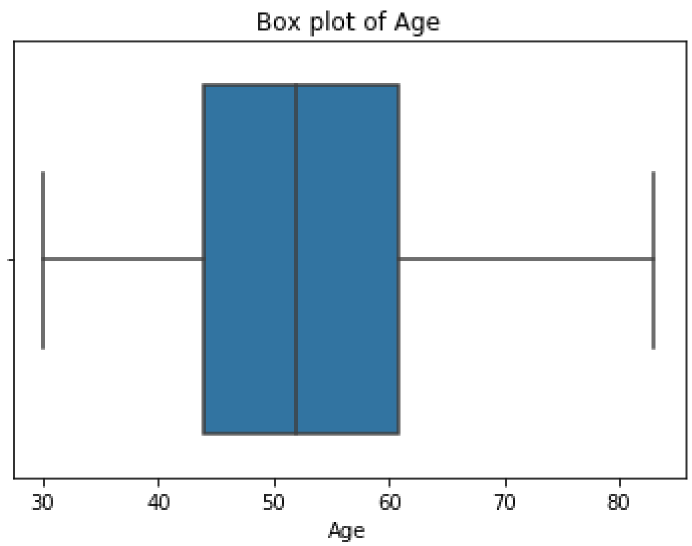
```
fig = sns.distplot(haber['age'])
plt.xlabel("Age")
plt.title("Dist plot of Age")
plt.show(fig)
```



Observation of Dist plot:

1. The age varies from 30 to around 82.
2. The mean of the age group is around 51.
3. Considering the density of the age we can consider this as a 'Balanced Data'.

In [8]:
```
#Box plot

fig = sns.boxplot(x='age',data=haber)
plt.xlabel("Age")
plt.title("Box plot of Age")
plt.show(fig)
```



Observations for Box plot:

1. Using the Box plot we can accurately say that, the minimum value is 30.
2. Considering the quartiles we can say the most of the people are around 45-62 years old.
3. Moreover from the Dist plot, Mean of the age group is clearly visible around 52.

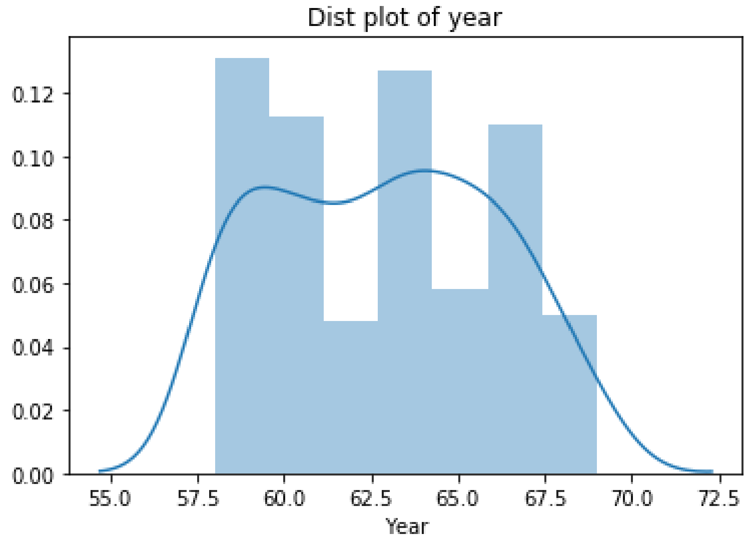
In [9]:
```
#Violin plot

fig = sns.violinplot(x='age',data=haber)
plt.xlabel("Age")
plt.title("Violin plot of Age")
plt.show(fig)
```

Observations for Violing plot:

1. As we know Violin plot is a mixture of both box plot and PDF, considering density we can say that most of the people of age group 45-60 undergone the surgery for cancer.

Univariate_Analysis of year

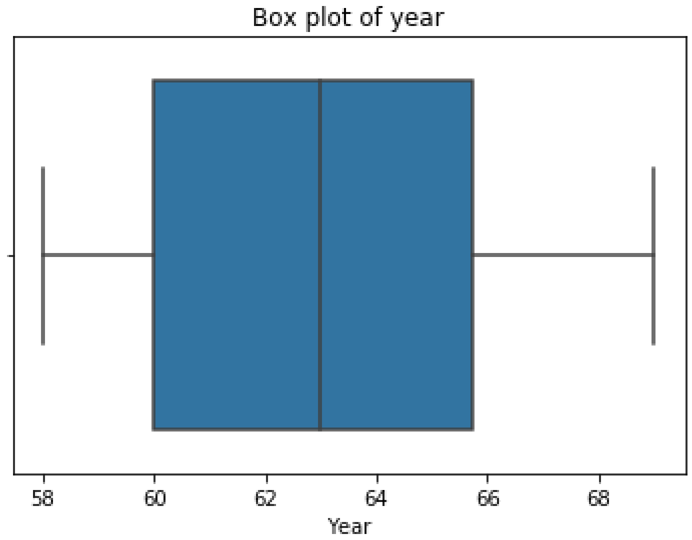
In [10]:
```
#Dist plot

fig = sns.distplot(haber['year'])
plt.xlabel("Year")
plt.title("Dist plot of year")
plt.show(fig)
```



Observation of Dist plot:

1. The Year in which patients undergone surgery varies from 58 to around 68.5.
2. The mean of the year is around 62.5.
3. Considering the density of the year we can consider this as a 'Balanced Data'.

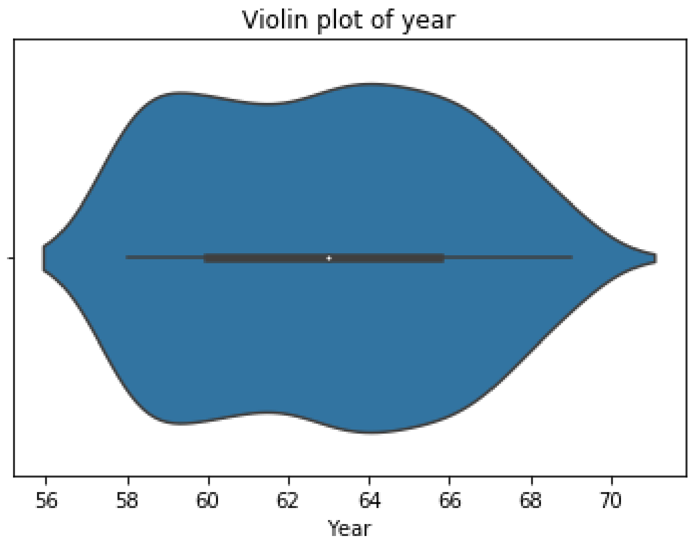
In [11]:
```
#Box plot

fig = sns.boxplot(x='year',data=haber)
plt.xlabel("Year")
plt.title("Box plot of year")
plt.show(fig)
```



Observations for Box plot:

1. Using the Box plot we can accurately say that, the minimum value is 60.
2. Considering the quartiles we can say the most of the people who have undergone surgery are around years 60-65.

In [12]:
```
#Violin plot

fig = sns.violinplot(x='year',data=haber)
plt.xlabel("Year")
plt.title("Violin plot of year")
plt.show(fig)
```

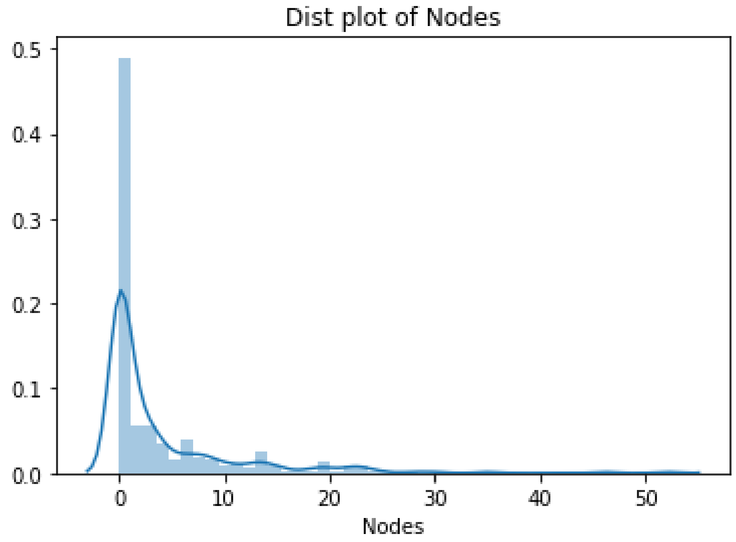


Observations for Violing plot:

1. Considering density we can say that most of the people who undergone the surgery for cancer from the year 58 to 69.

Univariate_Analysis of nodes

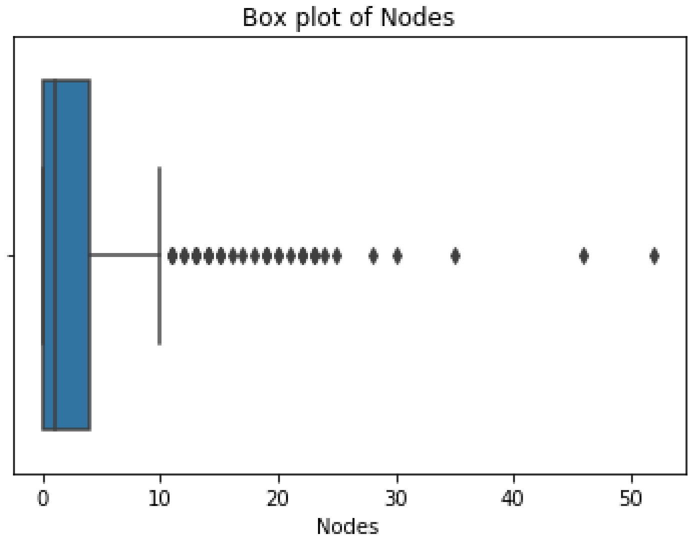
In [13]:
```
#Dist plot

fig = sns.distplot(haber['nodes'])
plt.xlabel("Nodes")
plt.title("Dist plot of Nodes")
plt.show(fig)
```



Observation of Dist plot:

1. The number of nodes varies from 0 to 52.
2. The mean of the nodes is around 2.5.
3. Considering the density of the nodes we can consider this as a 'UnBalanced Data'.

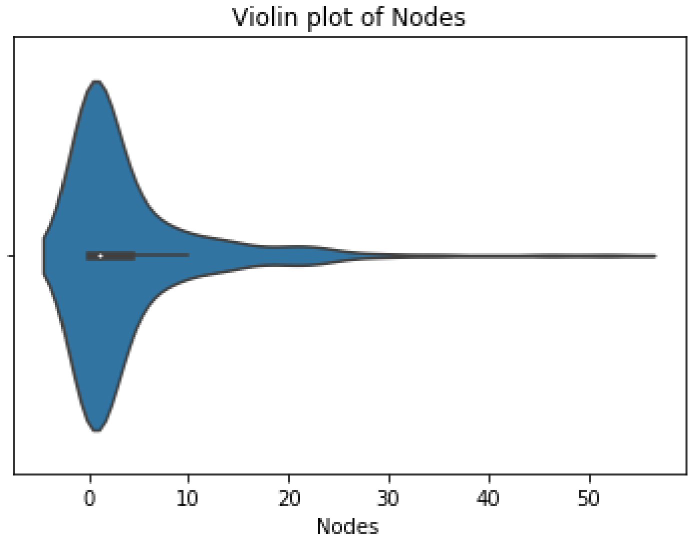
In [14]:
```
#Box plot

fig = sns.boxplot(x='nodes',data=haber)
plt.xlabel("Nodes")
plt.title("Box plot of Nodes")
plt.show(fig)
```



Observations for Box plot:

1. As the nodes data is a Unbalanced data we dont have a clear details about the nodes using the box plot.
2. Considering the quartiles we can say the mean is around 3.
3. The data consists of many Outliers out of Whiskers.

In [15]:
```
#Violin plot

fig = sns.violinplot(x='nodes',data=haber)
plt.xlabel("Nodes")
plt.title("Violin plot of Nodes")
plt.show(fig)
```

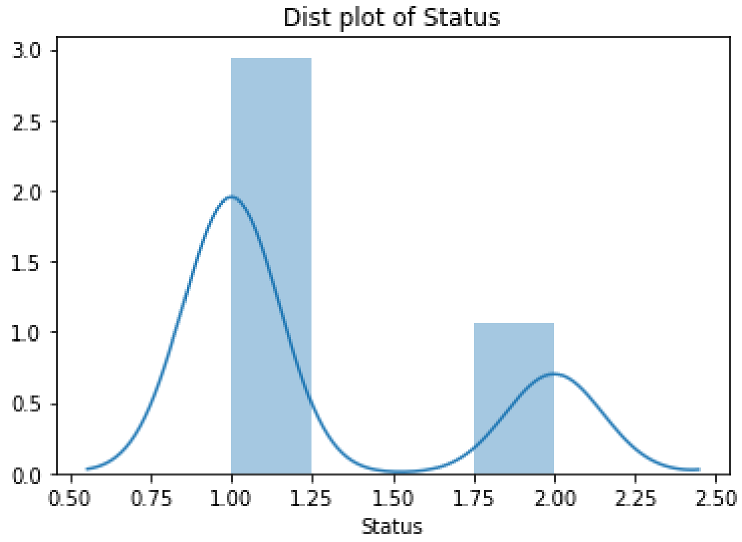


Observations for Violin plot:

1. Considering density we can say that most of the is not distributed propely.
2. We can say the density is huge around 2-3.

Univariate_Analysis of Status

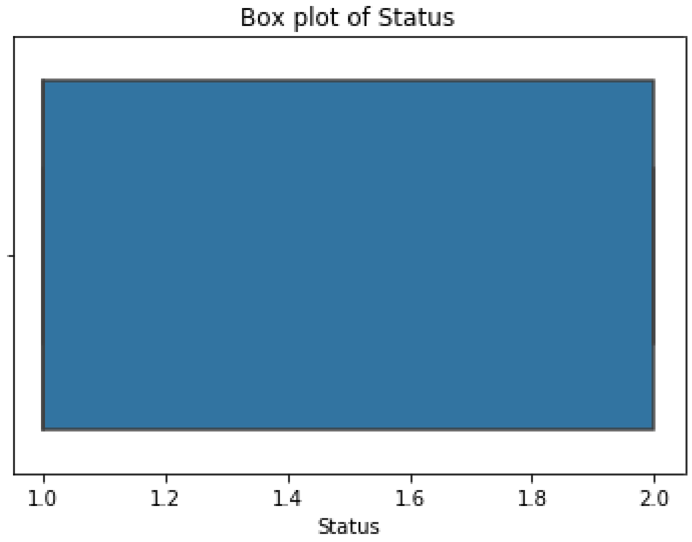
In [16]:
```
#Dist plot

fig = sns.distplot(haber['status'])
plt.xlabel("Status")
plt.title("Dist plot of Status")
plt.show(fig)
```



Observation of Dist plot:

1. The number of nodes varies from 1 to 2.
2. The mean of the group is around 1.5.
3. Considering the density of the status we can consider this as a 'UnBalanced Data'.

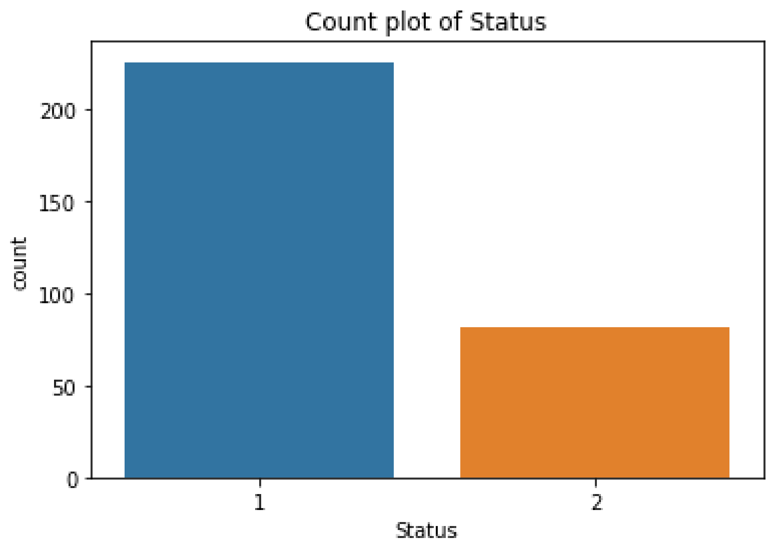
In [17]:
```
#Box plot

fig = sns.boxplot(x='status',data=haber)
plt.xlabel("Status")
plt.title("Box plot of Status")
plt.show(fig)
```



Observations for Box plot:

1. As we have only two different values there is no use of Box plot.

In [18]:
```
#count plot

fig = sns.countplot(data=haber,x='status')
plt.xlabel("Status")
plt.title("Count plot of Status")
plt.show(fig)
```



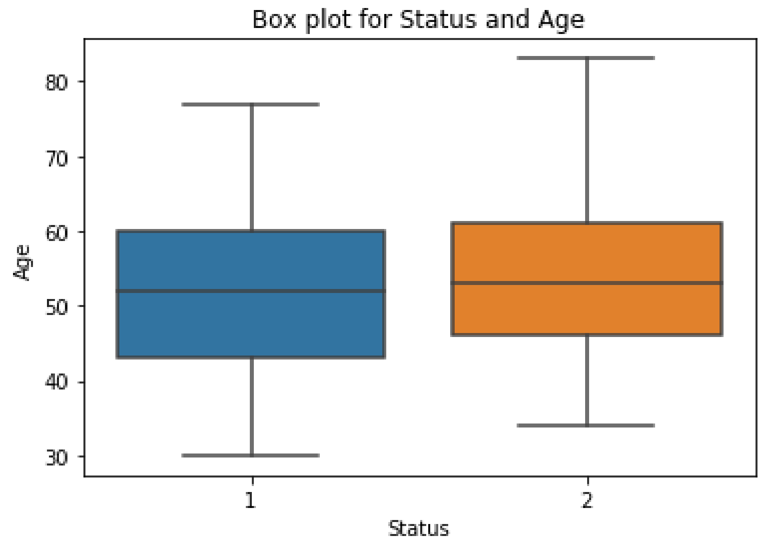
Observations for Count plot:

1. As we have only two different values 1 and 2 there is clear image stating that around 200+ people have survived more than 5+ years even after the cancer surgery.

Bivariate

Bivariate_Analysis of Status and Age

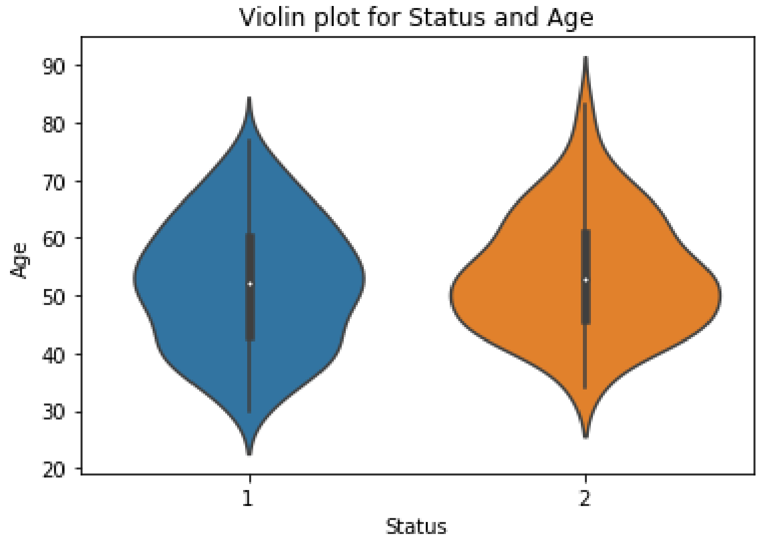
In [19]:
```
#Box plot

fig = sns.boxplot(x='status',y='age',data=haber)
plt.xlabel("Status")
plt.ylabel("Age")
plt.title("Box plot for Status and Age")
plt.show(fig)
```



Observations for Box Plot:

1. From above plot,we can see the mean age of people who survived more than 5 years is less than that of who survived less than 5 years.
2. As the length of two boxes is almost same, we can say that the dispersion of people who survived and who didnt survive is almost equal with the mean between 52-55.

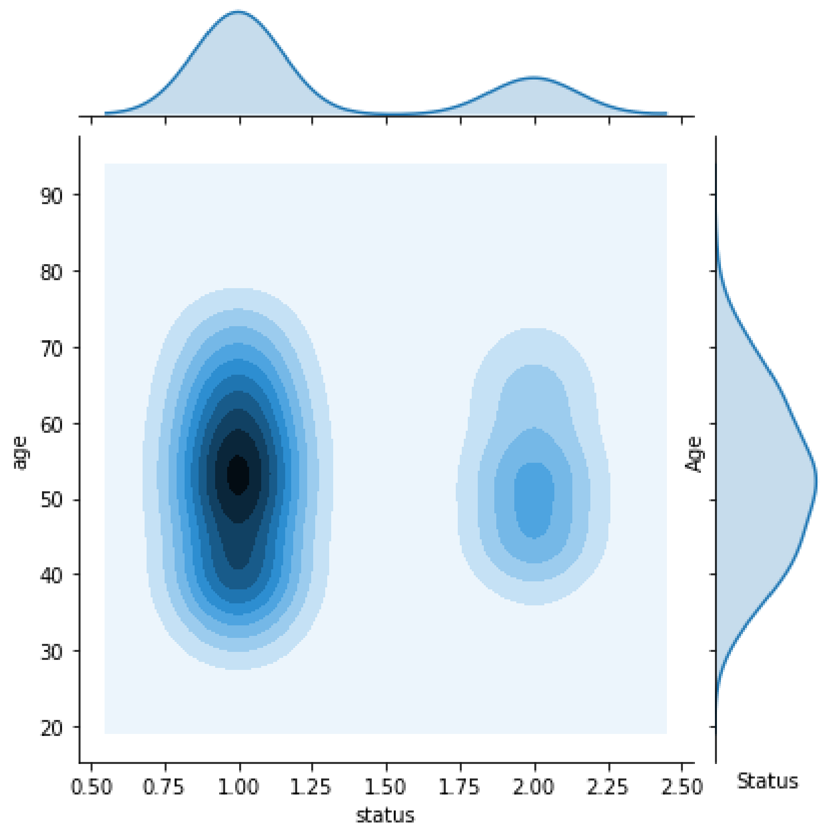
In [20]:
```
#Violin plot

fig = sns.violinplot(x='status',y='age',data=haber)
plt.xlabel("Status")
plt.ylabel("Age")
plt.title("Violin plot for Status and Age")
plt.show(fig)
```



Observations for Violin Plot:

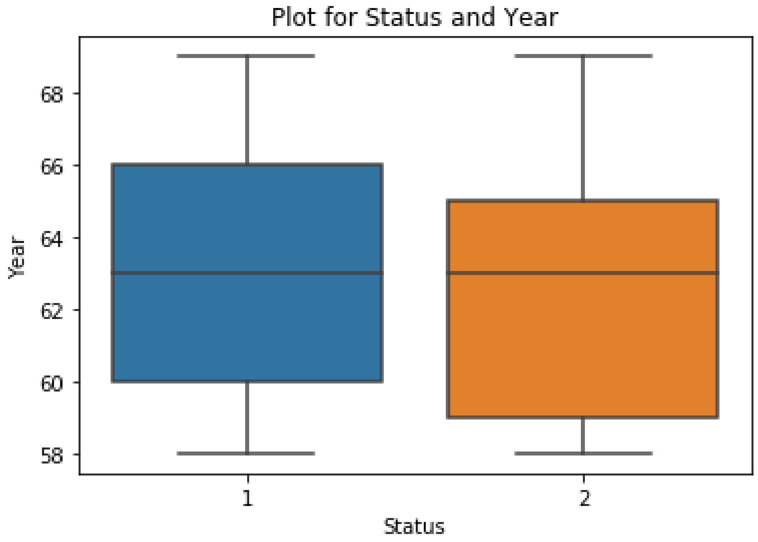
1. Comparing status and age we can say that, the density at the status=1 is equally distributed when compared to the denist =2.
2. The mean of the status who have survived >=5 and <5 is almost the same.
3. Most Importantly we can say that, People who have crossed the age of 50 has less chance to survive after the surgery.

In [21]:
```
#joint plot

fig = sns.jointplot(x="status", y="age", data=haber, kind="kde")
plt.xlabel("Status")
plt.ylabel("Age")
#plt.title("Joint plot for Status and Age")
plt.show(fig)
```



Bivariate_Analysis of status and Year

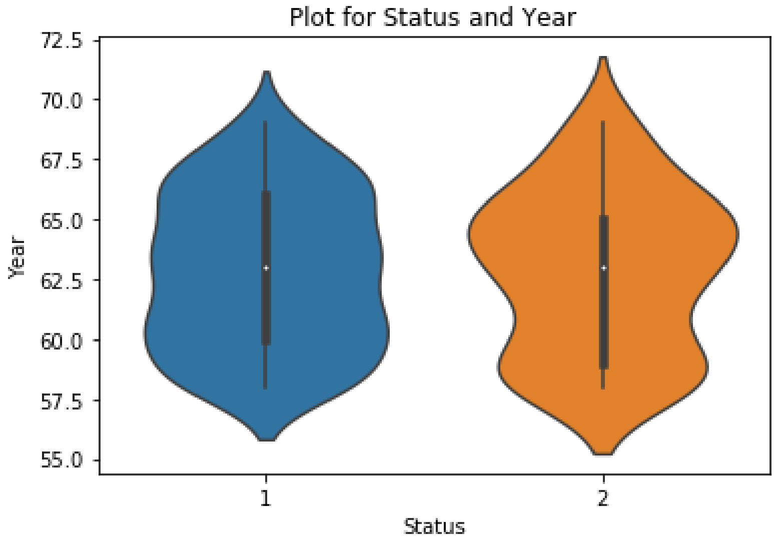
In [22]:
```
#Box plot

fig = sns.boxplot(x='status',y='year',data=haber)
plt.xlabel("Status")
plt.ylabel("Year")
plt.title("Plot for Status and Year")
plt.show(fig)
```



Observations for Box Plot:

1. Comparing status and year, the mean of the year is around 63.
2. People who have not survived more than 5 years starts from the year 59.

In [23]:
```
#Violin plot

fig = sns.violinplot(x='status',y='year',data=haber)
plt.xlabel("Status")
plt.ylabel("Year")
plt.title("Plot for Status and Year")
plt.show(fig)
```

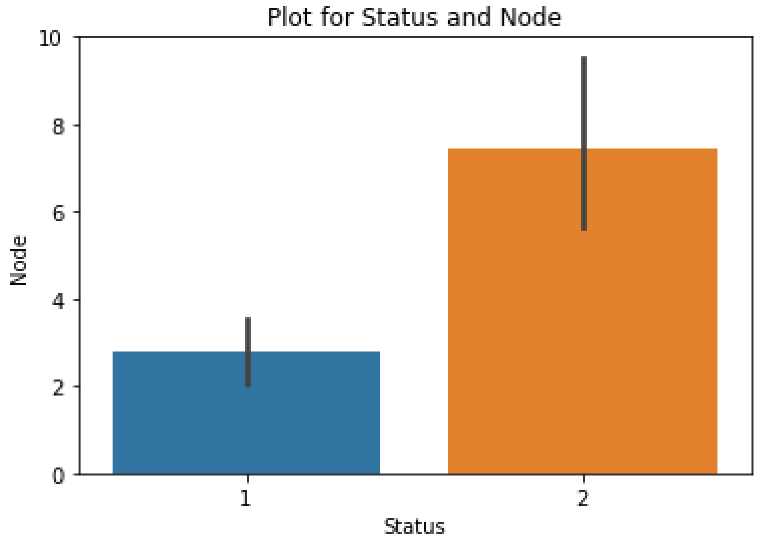


Observations for Violin Plot:

1. Comparing status and year we can say that, the density at the status=1 is equally distributed when compared to the denist =2.
2. The mean of the status who have survived >=5 and <5 is almost the same.

Bivariate_Analysis of status and Node

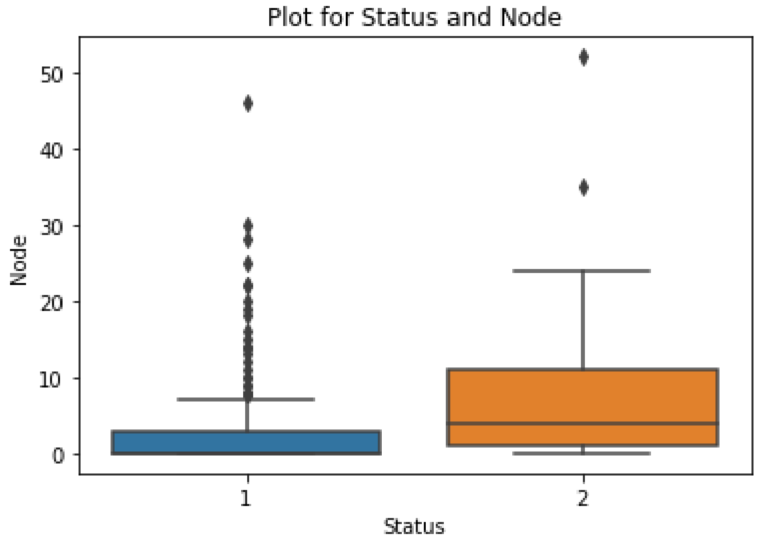
In [24]:
```
#Bar plot

fig = sns.barplot(x='status',y='nodes',data=haber)
plt.xlabel("Status")
plt.ylabel("Node")
plt.title("Plot for Status and Node")
plt.show(fig)
```



Observations for Bar plot:

1. People having less no. of nodes have the higher probability of living their life more than 5 years.

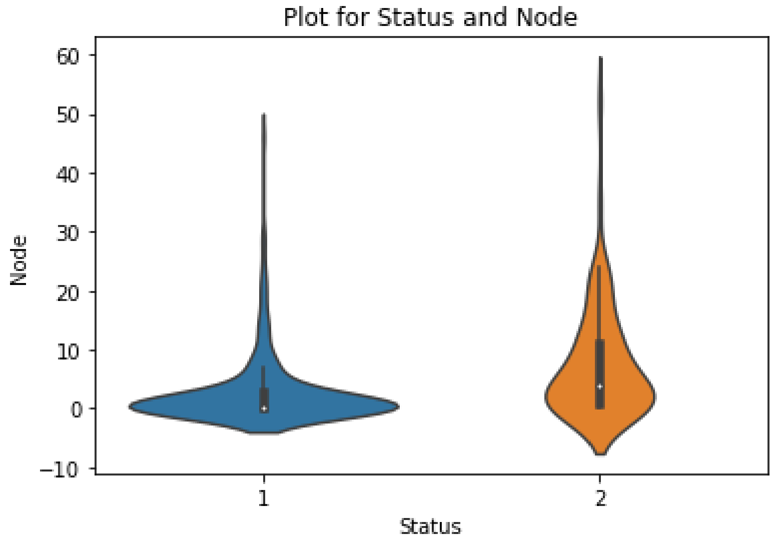
In [25]:
```
#Box plot

fig = sns.boxplot(x='status',y='nodes',data=haber)
plt.xlabel("Status")
plt.ylabel("Node")
plt.title("Plot for Status and Node")
plt.show(fig)
```



Observations for Box plot:

1. Here we observe that due to unbalanced distribution of data, there are many outliers formed even after the whiskers.
2. Comparitively the nodes are of high count for the people who didnt survive more than 5 year after the surgery.

In [26]:
```
#Violin plot

fig = sns.violinplot(x='status',y='nodes',data=haber)
plt.xlabel("Status")
plt.ylabel("Node")
plt.title("Plot for Status and Node")
plt.show(fig)
```

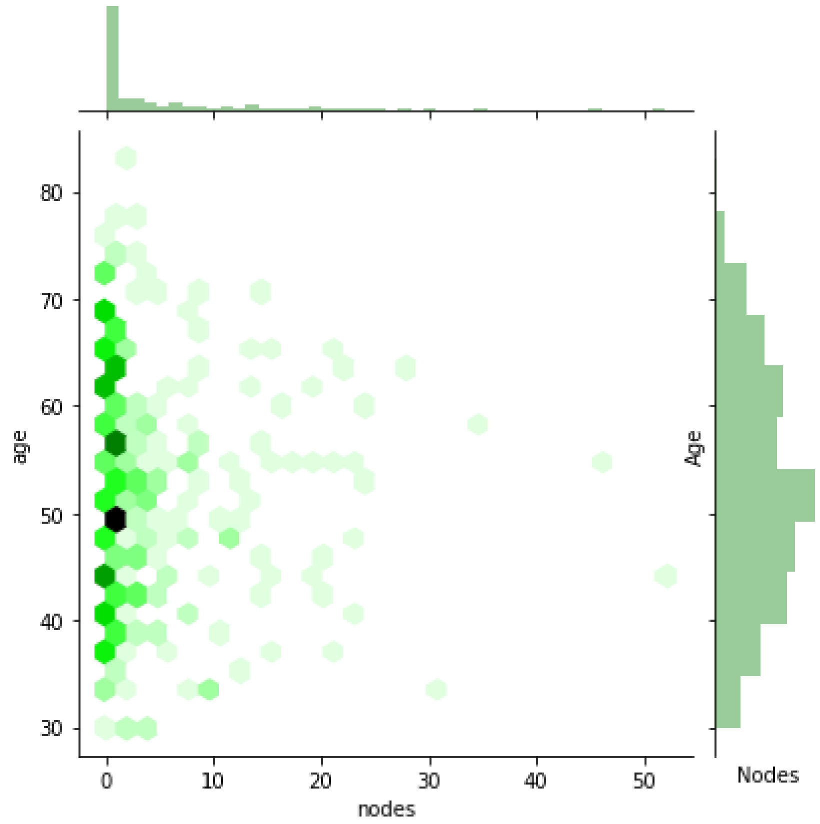


Observations for Violin plot:

1. People having less no.of nodes have the higher living probabilty.

Bivariate_Analysis of age and Nodes

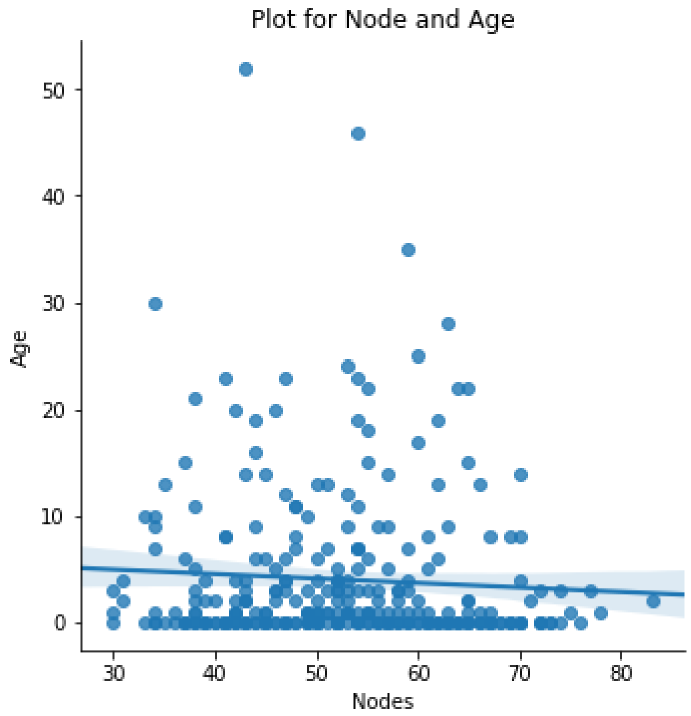
In [27]:
```
#Joint plot

fig = sns.jointplot(x='nodes',y='age',kind='hex',color='green',data=hab
er)
plt.xlabel("Nodes")
plt.ylabel("Age")
#plt.title("Plot for Node and Age")
plt.show(fig)
```



Observations for Joint plot:

1. The density is high between the age group 50 and 60 having nodes less than 5.

In [28]:
```
#lm plot

fig = sns.lmplot(x='age',y='nodes',data=haber)
plt.xlabel("Nodes")
plt.ylabel("Age")
plt.title("Plot for Node and Age")
plt.show(fig)
```



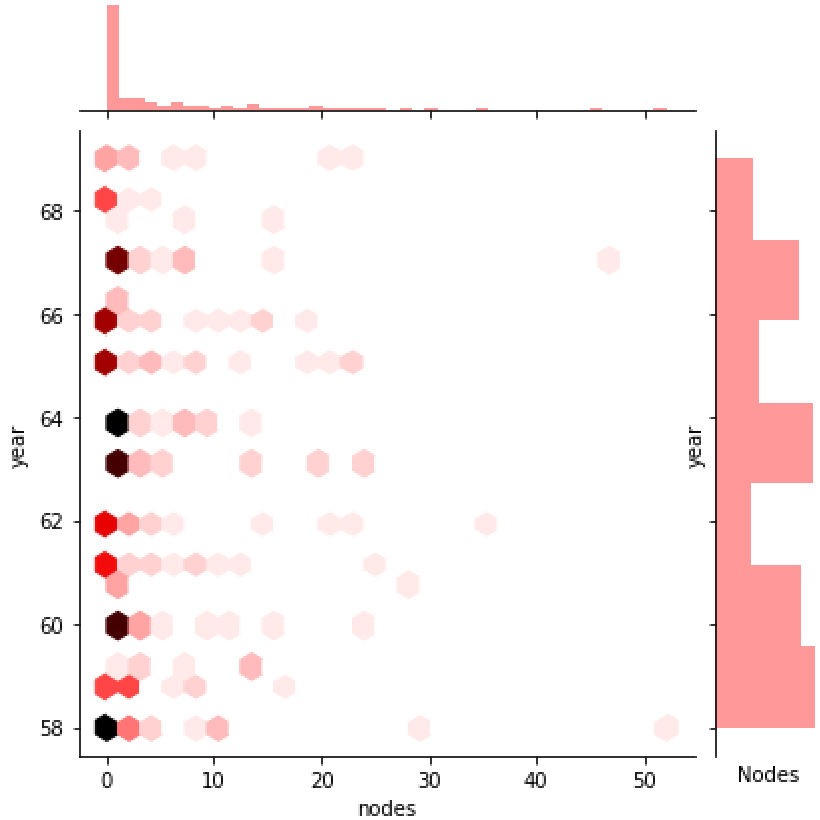
Observations for lm plot:

1. Only 4 people have 30+ nodes.

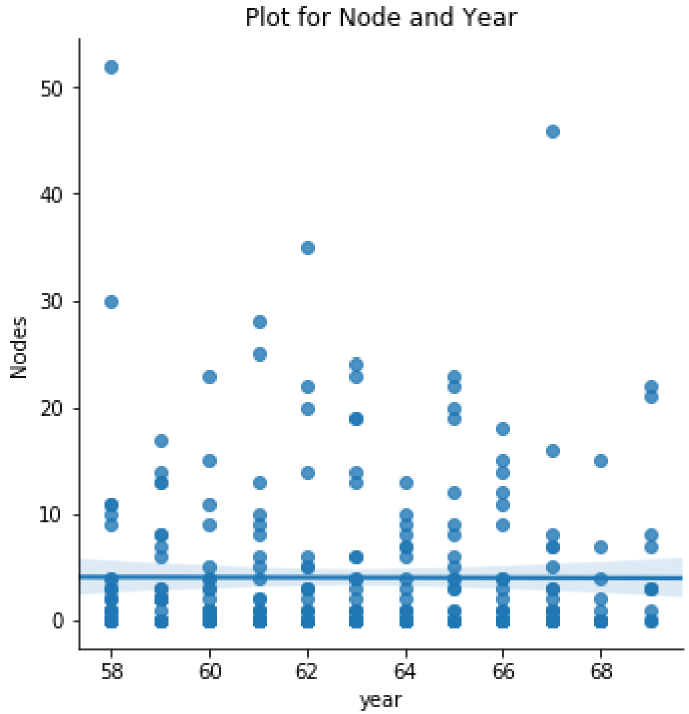
Bivariate_Analysis of year and Nodes

In [29]:
```
#Joint plot

fig = sns.jointplot(x='nodes',y='year',kind='hex',color='red',data=habe
r)
plt.xlabel("Nodes")
plt.ylabel("year")
```

```
#plt.title("Plot for Node and Year")
plt.show(fig)
```



In [30]:
```
#lm plot

fig = sns.lmplot(x='year',y='nodes',data=haber)
plt.xlabel("year")
plt.ylabel("Nodes")
plt.title("Plot for Node and Year")
plt.show(fig)
```

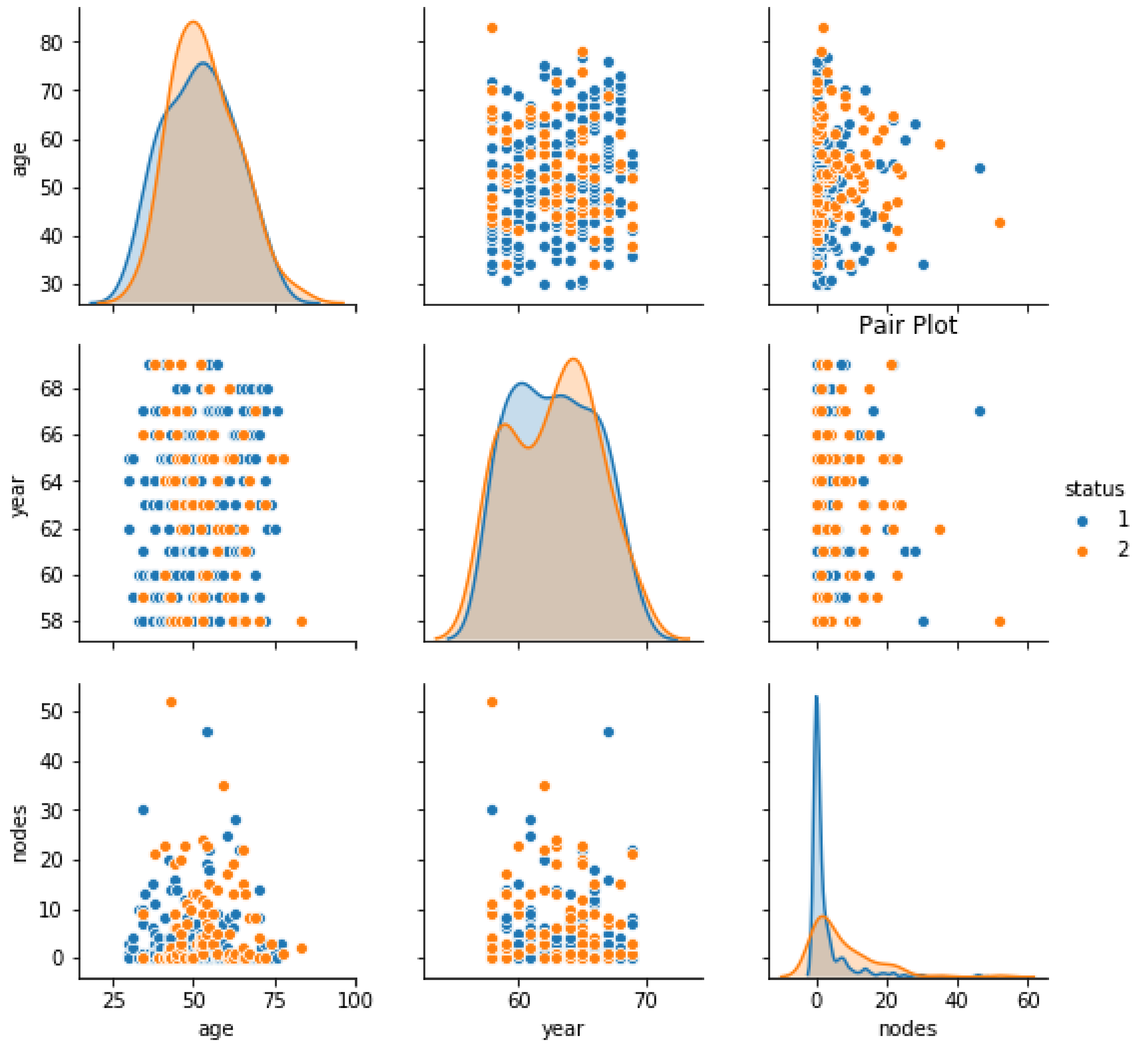


Observations for lm plot:

1. In the year 58, we have noticed 50+ nodes.
2. Appx after 9 years, at 67 we have observed a close call around 45 nodes.

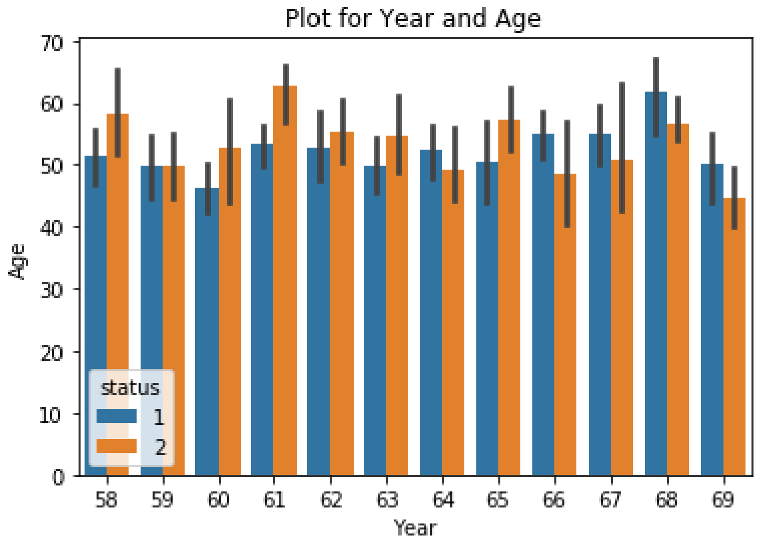
Multi variate

In [31]:
```
#pair plot

fig = sns.pairplot(data=haber,vars=['age','year','nodes'],hue='status')
plt.title("Pair Plot")
plt.show(fig)
```



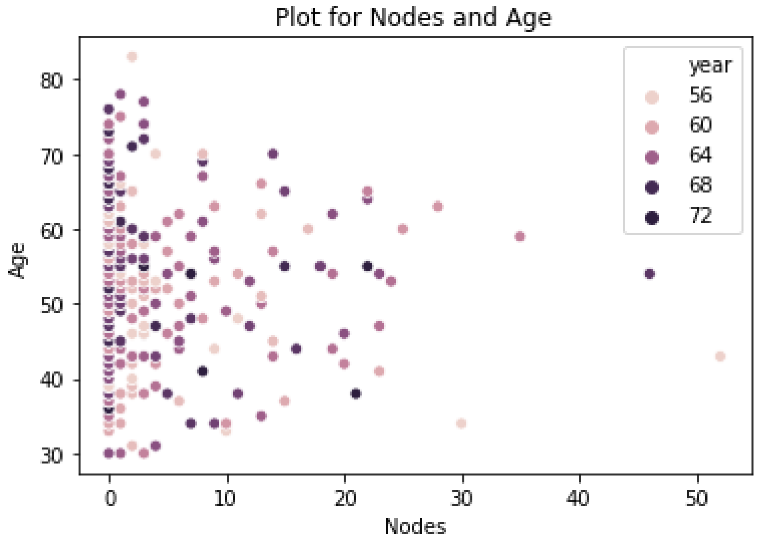
In [32]:

```
#Bar plot

fig = sns.barplot(x='year',y='age',hue='status',data=haber)
plt.xlabel("Year")
plt.ylabel("Age")
plt.title("Plot for Year and Age")
plt.show(fig)
```



In [33]:

```
#scatter plot

fig = sns.scatterplot(x='nodes',y='age',hue='year',color='black',data=h
aber)
plt.xlabel("Nodes")
plt.ylabel("Age")
plt.title("Plot for Nodes and Age")
plt.show(fig)
```



In [ ]:

Conclusion:

From the above plotting we can observe the following factors:

1. The data under the Age column is a Balanced data. Looks like to be a Gaussian distribution as the denisty of the data is distributed symmetrically on the both the side of curve.
2. From the Age data we can know that the Age varies from 30 to 82, with a mean of 52.
3. Considering the Violin plot of Age density we can say that most of the people of age group 45-60 undergone the surgery for cancer.
4. From the Nodes density we can say that the it 'Postive Skewed' as the tail is long towards on the positive ended direction. By this we also consider this data as a Unbalanced data.
5. Considering the the Class label Survival Status, stating that around 200+ people have survived more than 5+ years even after the cancer surgery.
6. Looking into the Bivariate Analysis of the Survival status and the Age, we understand that the Both the status people who have survived more than 5 years and also less than 5 are of the similar age group with a very minute difference between their respective means.

7. Taking into consideration of the year and the survival status using the box plot, I particularly observed that the patients who have undergone the treatment before 60, unfortunately have not survived more than 5 years.
8. Having less numbers nodes helped many of patients in surviving more than 5 years when compared to the patients with the more number of nodes.

In [ ]:

```

```