Assignment :-Exploratory Data Analysis

Objective:-We need to perform the exploratory data anaylisis to the habermans survival dataset and finally we need to tell the status of the patient i.e status=1,when patient survived 5 years or longer status-2,when patiet died within 5 years.

Dataset:-https://www.kaggle.com/gilsousa/habermans-survival-data-set

In [1]:
```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import warnings
warnings.filterwarnings("ignore")
hb=pd.read_csv('haberman.csv')#loading the file
hb.columns=['age','year','nodes','status']
#print(hb)#printing the dataset
#data set:https://www.kaggle.com/gilsousa/habermans-survival-data-set
```

In [18]:
```
#printing number of rows and columns
print(hb.shape)
```

```
(306, 4)
```

In [19]:
```
#printing number of attributes including class variable
print(hb.columns)
```

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

In [23]:
```
#data points per each class
#AS per the hebarman.csv file
# Survival status (class attribute) 1 = the patient survived 5 years or
 longer,2 = the patient died within 5 year
hb['status'].value_counts()
```

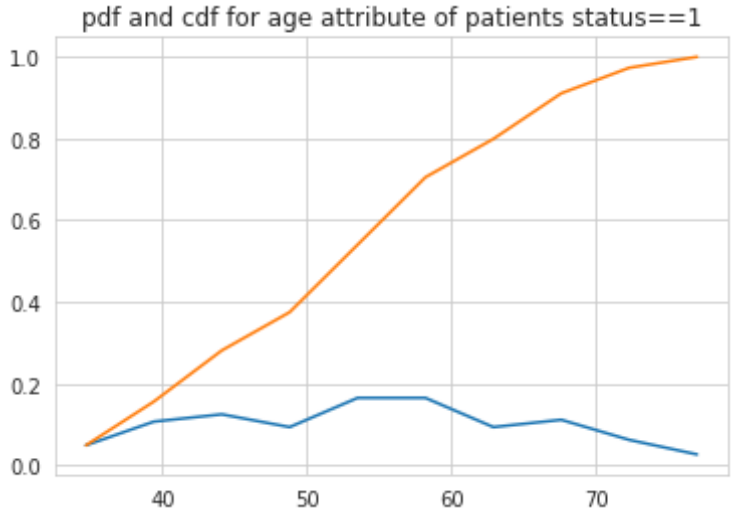
Out[23]:
```
1    225
2     81
Name: status, dtype: int64
```

Observations:- 1)The dataset is having total 306 rows and 4 columns in which 3 columns are age,year,nodes and status is 4th column which is class variable. 2) if status==1 then patient will surive 5 or more years if statys==2 then patient will die less than 5 years 3) total 306 patients 225 patients will belongs to class of status=1 and 81 patients will belongs to class of status=2

Univariate Analysis

In [9]:
```
hb_1=hb.loc[hb["status"]==1];#loading all details of patients whose bel
ongs to class of status==1 into hb_1
hb_2=hb.loc[hb["status"]==2];#loading all details of patients whose bel
ongs to class of status==2 into hb_2
#age of patients status=1
counts1,bin_edges1=np.histogram(hb_1['age'],bins=10,density=True)
pdf=counts1/(sum(counts1))
print(pdf)
print(bin_edges1)
cdf=np.cumsum(pdf)
plt.plot(bin_edges1[1:],pdf)
plt.plot(bin_edges1[1:], cdf)
plt.title("pdf and cdf for age attribute of patients status==1")
plt.show();
```

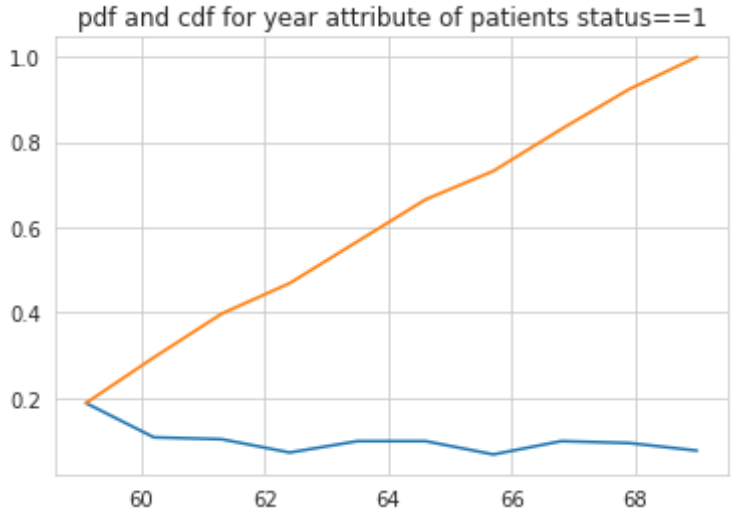
```
[0.04910714 0.10714286 0.125      0.09375    0.16517857 0.16517857
 0.09375    0.11160714 0.0625     0.02678571]
[30.  34.7 39.4 44.1 48.8 53.5 58.2 62.9 67.6 72.3 77. ]
```



Observation:-From the above graph we can say that there are exactly 40% of patients whose age is <=50 years and the surveillance status =1. The reason why we can say exactly because exactly cdf curve is touching at point (50,0.4) on x,y axis respectively.

In [10]:
```
#year of patients status=1
counts1,bin_edges1=np.histogram(hb_1['year'],bins=10,density=True)
pdf=counts1/(sum(counts1))
print(pdf)
print(bin_edges1)
cdf=np.cumsum(pdf)
plt.plot(bin_edges1[1:],pdf)
plt.plot(bin_edges1[1:], cdf)
plt.title("pdf and cdf for year attribute of patients status==1")
plt.show();
```

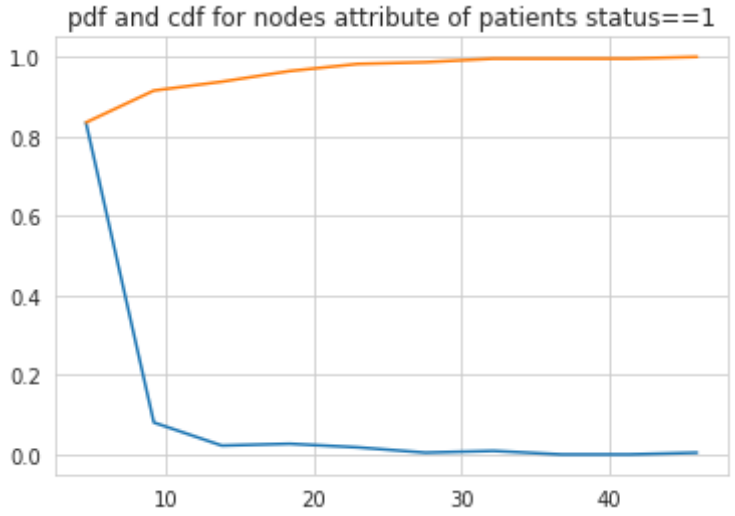
```
[0.1875     0.10714286 0.10267857 0.07142857 0.09821429 0.09821429
 0.06696429 0.09821429 0.09375    0.07589286]
[58.  59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
```



Observation:- From the above graph we can say that there are exactly 60% of patients whose born in the years <=1964 and surveillance status=1

In [11]:
```
#nodes of patients status=1
counts1,bin_edges1=np.histogram(hb_1['nodes'],bins=10,density=True)
pdf=counts1/(sum(counts1))
print(pdf)
print(bin_edges1)
cdf=np.cumsum(pdf)
plt.plot(bin_edges1[1:],pdf)
plt.plot(bin_edges1[1:], cdf)
plt.title("pdf and cdf for nodes attribute of patients status==1")
plt.show();
```

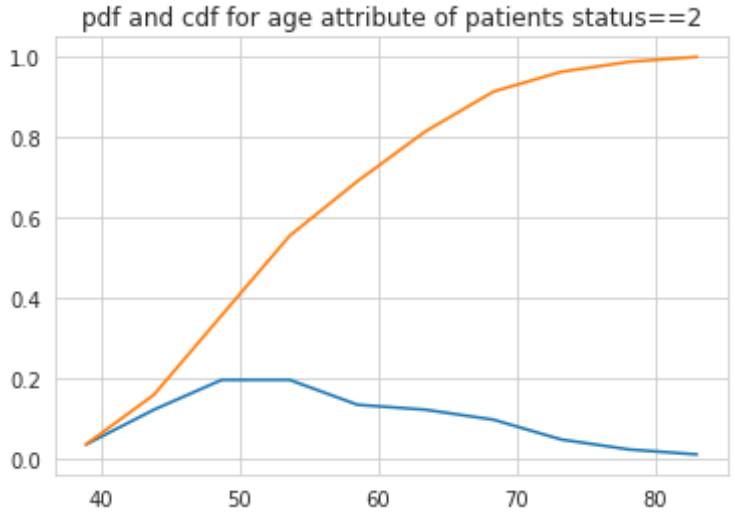
```
[0.83482143 0.08035714 0.02232143 0.02678571 0.01785714 0.00446429
 0.00892857 0.         0.         0.00446429]
[ 0.   4.6  9.2 13.8 18.4 23.  27.6 32.2 36.8 41.4 46. ]
```



Observation:- From the above graph we can say that there are approximately 91% of patients whose postively detected auxilary nodes are <=10 and surveillance status=1

In [12]:
```
hb_1=hb.loc[hb["status"]==1];#loading all details of patients whose bel
ongs to class of status==1 into hb_1
hb_2=hb.loc[hb["status"]==2];#loading all details of patients whose bel
ongs to class of status==2 into hb_2
#age of patients status=2
counts1,bin_edges1=np.histogram(hb_2['age'],bins=10,density=True)
pdf=counts1/(sum(counts1))
print(pdf)
print(bin_edges1)
cdf=np.cumsum(pdf)
plt.plot(bin_edges1[1:],pdf)
plt.plot(bin_edges1[1:], cdf)
plt.title("pdf and cdf for age attribute of patients status==2")
plt.show();
```

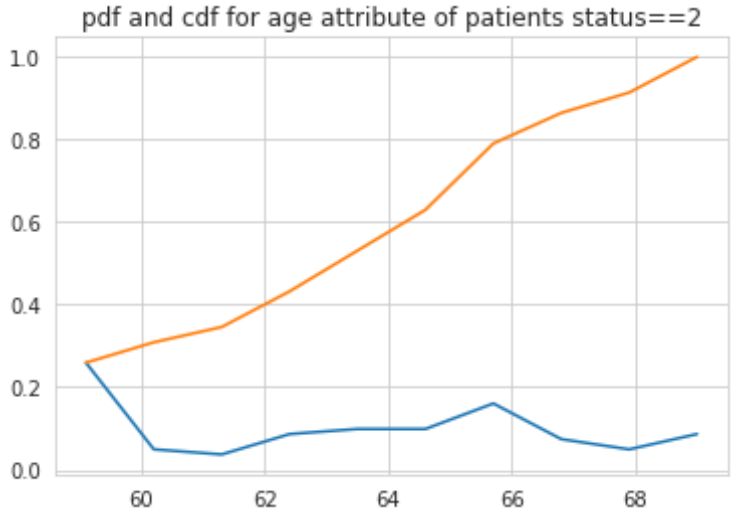
```
[0.03703704 0.12345679 0.19753086 0.19753086 0.13580247 0.12345679
 0.09876543 0.04938272 0.02469136 0.01234568]
[34.  38.9 43.8 48.7 53.6 58.5 63.4 68.3 73.2 78.1 83. ]
```



Observation:- From the above graph we can say that there are exactly 40% of patients whose age is <=50 years and the surveillance status =2.

In [14]:
```
hb_1=hb.loc[hb["status"]==1];#loading all details of patients whose bel
ongs to class of status==1 into hb_1
hb_2=hb.loc[hb["status"]==2];#loading all details of patients whose bel
ongs to class of status==2 into hb_2
#year of patients status=2
counts1,bin_edges1=np.histogram(hb_2['year'],bins=10,density=True)
pdf=counts1/(sum(counts1))
print(pdf)
print(bin_edges1)
cdf=np.cumsum(pdf)
plt.plot(bin_edges1[1:],pdf)
plt.plot(bin_edges1[1:], cdf)
plt.title("pdf and cdf for age attribute of patients status==2")
plt.show();
```

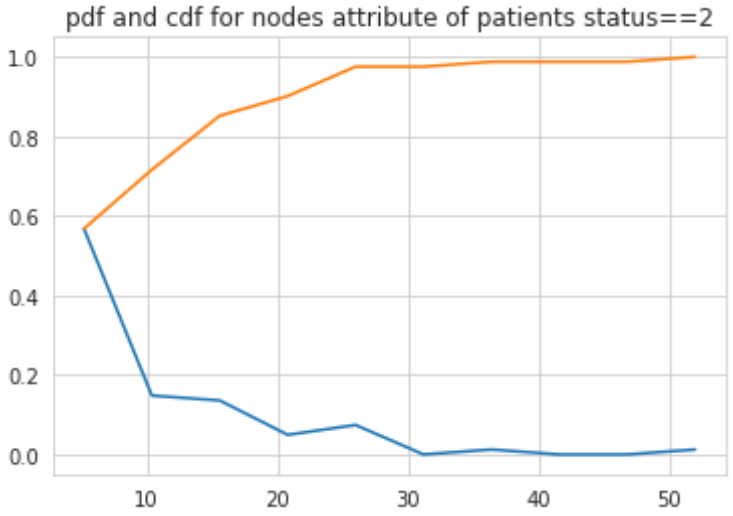
```
[0.25925926 0.04938272 0.03703704 0.08641975 0.09876543 0.09876543
 0.16049383 0.07407407 0.04938272 0.08641975]
[58.  59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
```



Observation:- From the above graph we can say that there are exactly 80% of patients whose born in the years <=1966 and surveillance status=2

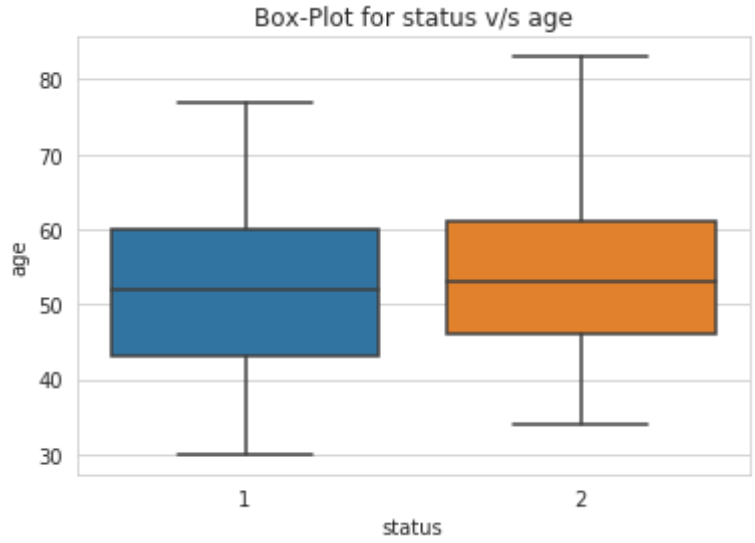
In [15]:
```
#nodes of patients status=2
counts1,bin_edges1=np.histogram(hb_2['nodes'],bins=10,density=True)
pdf=counts1/(sum(counts1))
print(pdf)
print(bin_edges1)
cdf=np.cumsum(pdf)
plt.plot(bin_edges1[1:],pdf)
plt.plot(bin_edges1[1:], cdf)
plt.title("pdf and cdf for nodes attribute of patients status==2")
plt.show();
```

```
[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. ]
```



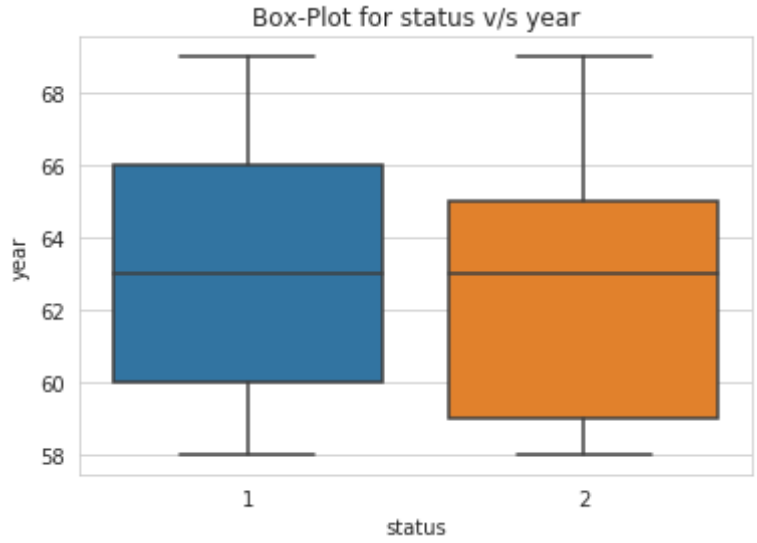
Observation:- From the above graph we can say that there are approximately 90% of patients whose postively detected auxilary nodes are <=20 and surveillance status=2

In [16]:
```
#Univariage Analysis
#Box plot
sns.set_style("whitegrid")
sns.boxplot(x='status',y='age',data=hb)
plt.title("Box-Plot for status v/s age")
plt.show()
```



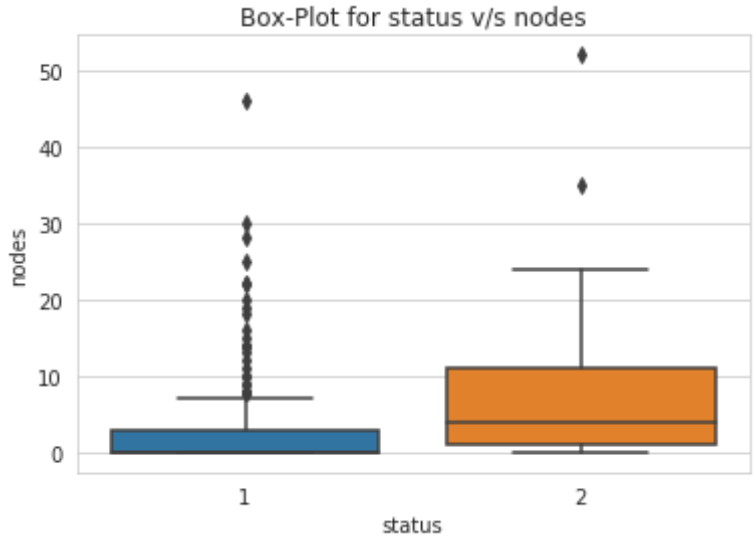
Observations:- 1)if the patient age is lies in between 30 and 31 will belongs to class of status=1 2)if the patient age is >60 will belongs to class of status=2 3)the patient whose status=1 and 75% of such pateints are having age in between 42 and 60

In [17]:
```
sns.set_style("whitegrid")
sns.boxplot(x='status',y='year',data=hb)
plt.title("Box-Plot for status v/s year")
plt.show()
```



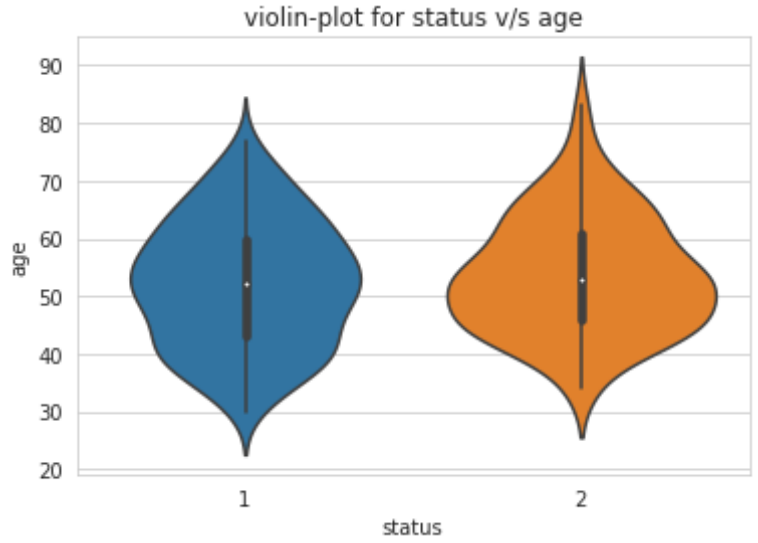
Observation:- f the patient was born on the year less than 1960 will definitely will belongs to class of status=2 and if the patient was born on the year greater than 1965 will definitely will belongs to class of status=1 and in between we cannot say because overlapping is coming so exactly we cannot design any model if the year is lies in between like 1960 and 1964

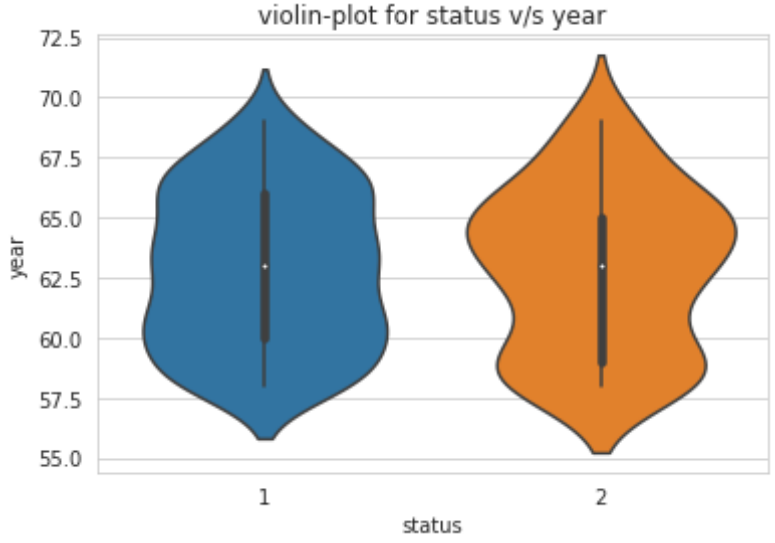
In [18]:
```
#Univariage Analysis
#Box plot
sns.set_style("whitegrid")
sns.boxplot(x='status',y='nodes',data=hb)
plt.title("Box-Plot for status v/s auxilary nodes")
plt.show()
```

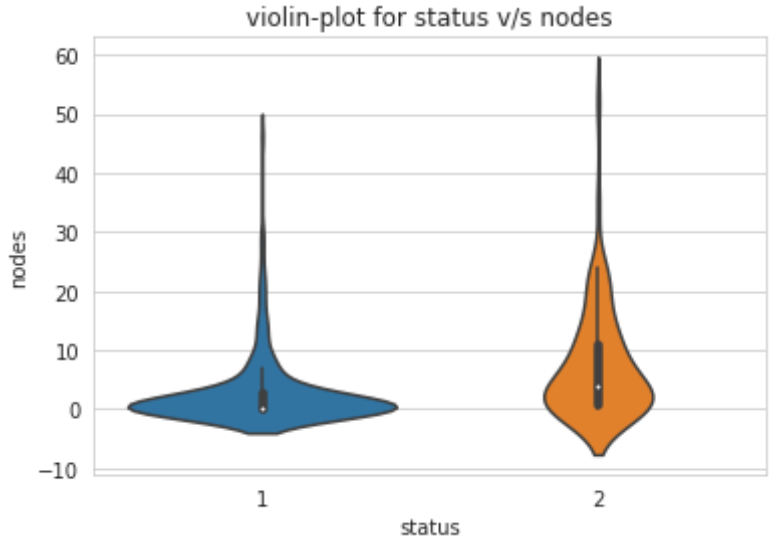


Observation:- the positively detected auxilary nodes will not me much helpfull for designing any model

In [6]:
```
import warnings
warnings.filterwarnings("ignore")
sns.set_style("whitegrid")
sns.violinplot(x="status", y="age",data=hb,size=6)
plt.title("violin-plot for status v/s age")
plt.show()
sns.violinplot(x="status", y="year",data=hb,size=6)
plt.title("violin-plot for status v/s year")
plt.show()
sns.violinplot(x="status", y="nodes",data=hb,size=6)
plt.title("violin-plot for status v/s nodes")
plt.show()
```







Observation from 3 violin plots:-

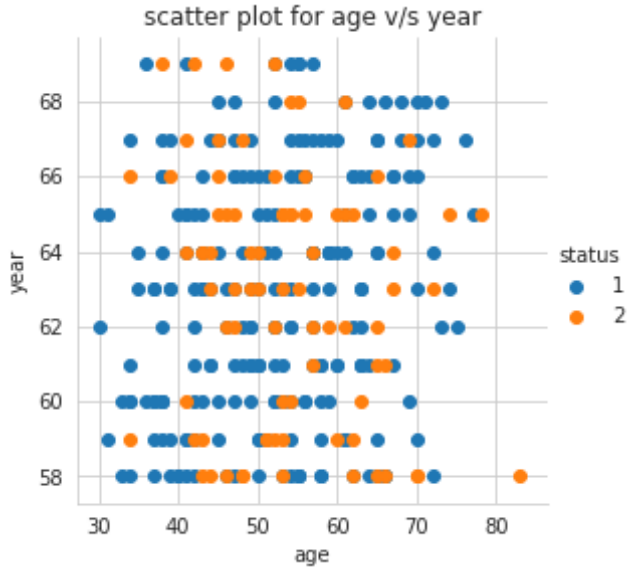
from violin plots we will get information of boxplots as well as pdf of data.

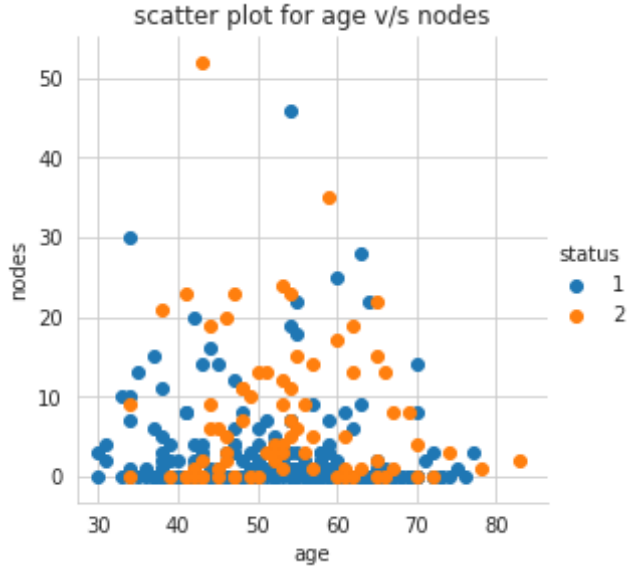
from that pdf we can say whether it follows some standard distributions or not if you see the age values of patients whose status=1 will follow sligthly normal distribution and remaining attributes will not follow any distributions.

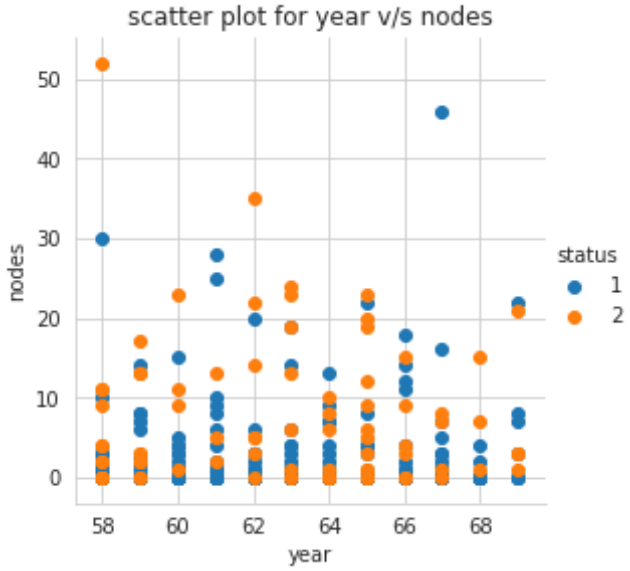
Bi-variate Analysis

In [7]:
```
sns.set_style("whitegrid");
sns.FacetGrid(hb,hue="status",size=4).map(plt.scatter,"age","year").add
_legend();
plt.title("scatter plot for age v/s year")
plt.show();
sns.FacetGrid(hb,hue="status",size=4).map(plt.scatter,"age","nodes").ad
d_legend();
plt.title("scatter plot for age v/s nodes")
plt.show();
```

```
sns.FacetGrid(hb,hue="status",size=4).map(plt.scatter,"year","nodes").a
dd_legend();
plt.title("scatter plot for year v/s nodes")
plt.show();
```

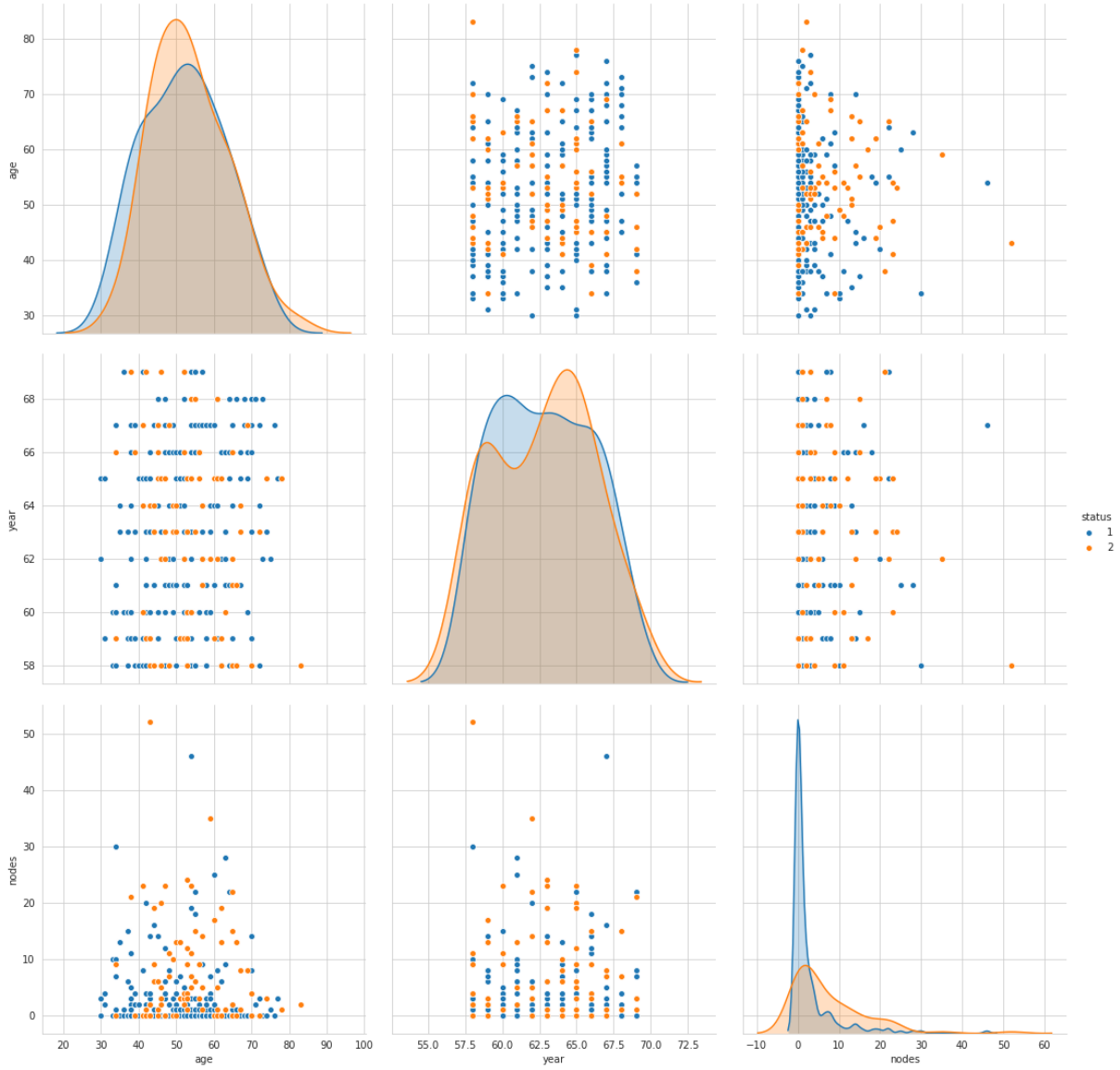






Observation:- From the above scatter plots graphs we cannot derive much information because data is fully overlapped.

In [52]:
```
#pairplot
sns.set_style("whitegrid")
plt.close()
sns.pairplot(hb,hue='status',vars=['age','year','nodes'],size=5)
plt.show()
```



Observation:- From the above pair plots we cannot conclude much information like which two features are most usefull to identify the status of the patient because the data is fully overlapped.

Conclusion From Bivariate analysis:-

1)Generally Bi-variate analysis we will do because of two or more features combinely will help for identifying the class 2)In this bivariate analysis will not be helpfull much because from either pair plots or scatter plots we didn't get much information to identify the patient will belongs to which class because of data is having fully overlapped nature so in this case Bi-Variate analysis will not be much helpful.