Firstly we will import all the required python libraries to analyise haberman dataset

In [5]:

```
import warnings
warnings.filterwarnings('ignore')
```

In [10]:

```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
```

Now we are going to find the number of rows and columns in haberman dataset

In [15]:

```
haberman=pd.read_csv('haberman.csv')
print(haberman.shape)
```

(306, 4)

Now we are going to find the Columns name for this dataset or the features

In [16]:

```
print(haberman.columns)
```

Index(['age', 'operation_year', 'axil_nodes', 'status'], dtype='object')

Now we will calculate the total counts for different status

In [17]:

```
haberman['status'].value_counts()
```

Out[17]:

1 225
2 81
Name: status, dtype: int64

So here we noticed that we have 2 types of status 1 and 2 and the total count for status 1 are 225 and for status 2 are 81

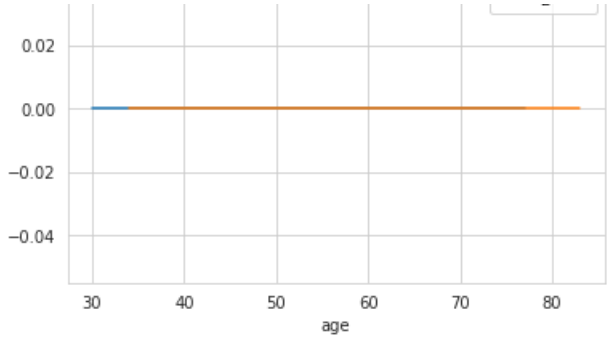
objective:

Here our objective is to classify a new data according to the existing classes.
For eg- If a got a new patient details and we have to check his survival status.

UNIVARIATE ANALYSIS

In [39]:

```
haberman_1=haberman.loc[haberman['status']==1];
haberman_2=haberman.loc[haberman['status']==2];
plt.plot(haberman_1['age'],np.zeros_like(haberman_1['age']),label='status\n''1')
plt.plot(haberman_2['age'],np.zeros_like(haberman_2['age']),label='2')
plt.title('One-dimensional plot for age')
plt.xlabel('age')
plt.legend()
plt.show()
```



Conclusion-From above figure we observed that the people whose age is less than 35 are in true or living condtion.

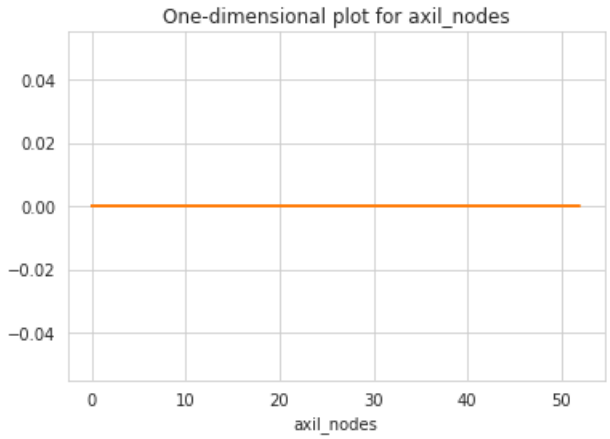
In [41]:

```
haberman_1=haberman.loc[haberman['status']==1];
haberman_2=haberman.loc[haberman['status']==2];
plt.plot(haberman_1['operation_year'],np.zeros_like(haberman_1['operation_year']),label='status\n"1')
plt.plot(haberman_2['operation_year'],np.zeros_like(haberman_2['operation_year']),label='2')
plt.title('One-dimensional plot for operation-year')
plt.xlabel('operation_year')
plt.show()
```

conclusion- operation_year variable is not that much relevent to analyise the survival status.

In [42]:

```
haberman_1=haberman.loc[haberman['status']==1];
haberman_2=haberman.loc[haberman['status']==2];
plt.plot(haberman_1['axil_nodes'],np.zeros_like(haberman_1['axil_nodes']),label='status\n"1')
plt.plot(haberman_2['axil_nodes'],np.zeros_like(haberman_2['axil_nodes']),label='2')
plt.title('One-dimensional plot for axil_nodes')
plt.xlabel('axil_nodes')
plt.show()
```



Conclusion- Axil_nodes variable is also not that much relevent to analyise survival status.

from the above plots we conclude that age would be relevent variable to analyise the data points

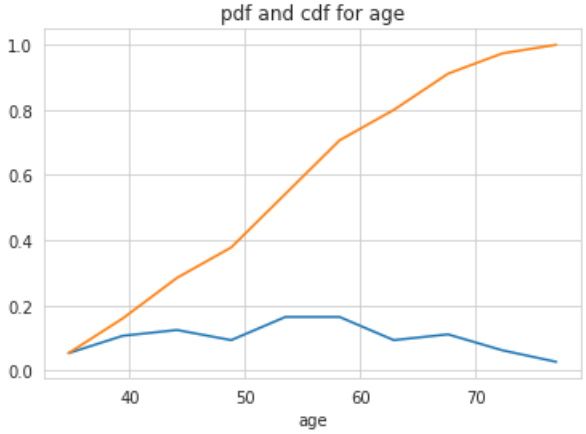
from the above plots we conclude that age would be relevent variable to analyse the data points

PDF AND CDF

In [44]:

```
total,bin_edges=np.histogram(haberman_1['age'],bins=10,density=True)
pdf=total/sum(total)
print(pdf)
print(bin_edges)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.title('pdf and cdf for age')
plt.xlabel('age')
plt.show()
```

```
[0.05333333 0.10666667 0.12444444 0.09333333 0.16444444 0.16444444
 0.09333333 0.11111111 0.06222222 0.02666667]
[30.  34.7 39.4 44.1 48.8 53.5 58.2 62.9 67.6 72.3 77. ]
```

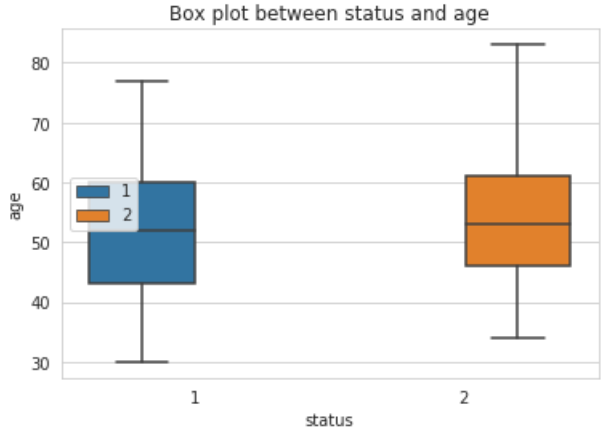


Conclusion- People whose age are less than 35 will survive.

BOX PLOT

In [55]:

```
sns.boxplot(hue='status',x='status',y='age',data=haberman)
plt.title('Box plot between status and age')
plt.legend()
plt.show()
```



VIOLIN PLOT

In [57]:

```
sns.violinplot(x='status',y='age',hue='status',data=haberman)
plt.title('violin plot between status and age')
```

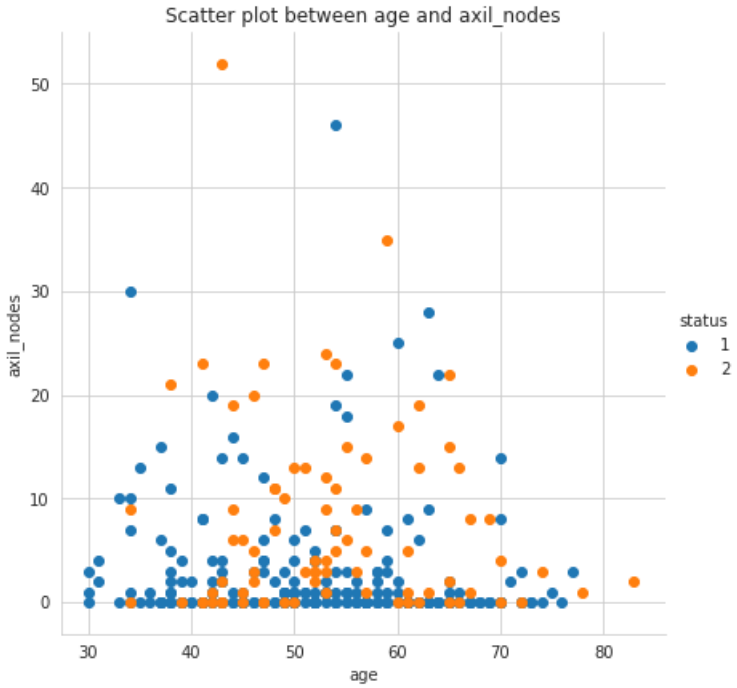
```
plt.show()
```

BIVARIATE ANALYSIS

SCATTER PLOT

In [30]:

```
sns.set_style('whitegrid');
sns.FacetGrid(haberman,hue='status',height=6).map(plt.scatter,'age','axil_nodes').add_legend();
plt.title('Scatter plot between age and axil_nodes')
plt.show();
```

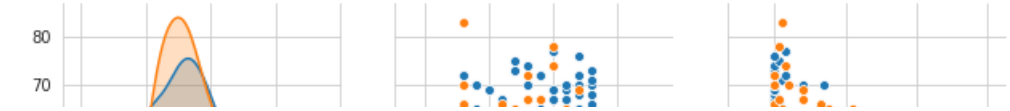


Conclusion- Quite difficult to analyise as these points are mixed up and are inseperable but the number of people died are less than living as points with 1 status are more.

PAIR PLOT

In [29]:

```
sns.set_style('whitegrid');
sns.pairplot(haberman,hue='status',vars=['age','operation_year','axil_nodes'],height=3);
plt.show()
```



OBSERVATIONS

Here in my first assignment i have taken haberman dataset and i have performed following operations- 1)Found number of rows and columns. 2)Found the Column names. 3)Calculate total counts for different status values. 4)Written the objective of this analysis. 5)found the best variable out of three by univariate analysis. 6)Performed PDF and CDF. 7)Box Plot 8)Violin Plot 9)Bivariate analysis with pair plot and scatter plot.

CONCLUSION

1)As for 2 different status we got total count as 225 and 81 which makes it unbalanced dataset. 2)I didnt get much information from pair plot.