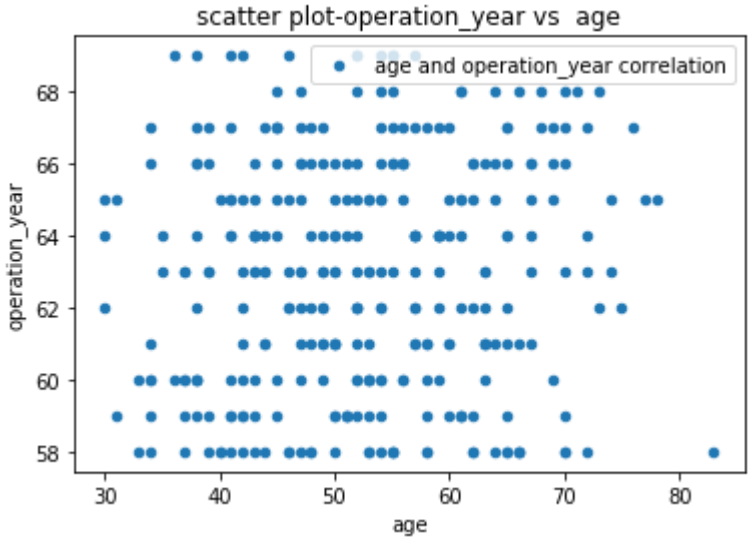
In [1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

haberman=pd.read_csv("haberman.csv")
print (haberman.shape)


#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 =


#Objective: find coorelation if any on the survival status based on age,axil_nodes

print (haberman.columns)
```

```
(306, 4)
Index(['age', 'operation_year', 'axil_nodes', 'surv_status'], dtype='object')
```

In [2]:
```
haberman["surv_status"].value_counts()
```

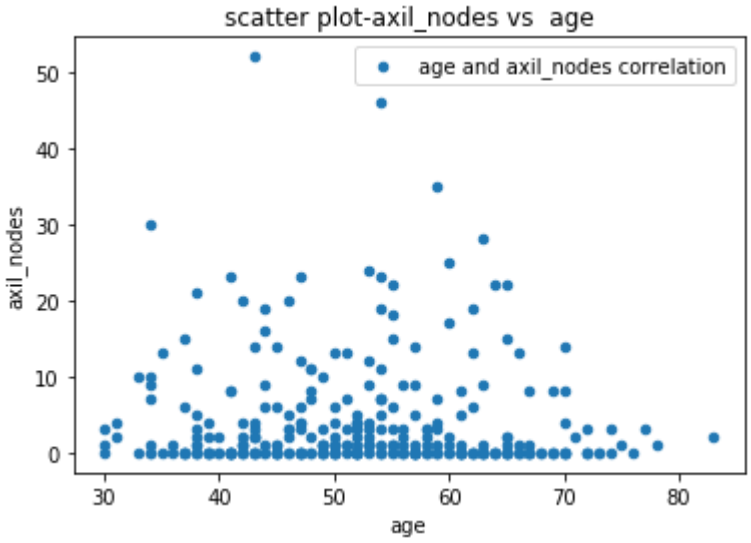
Out[2]:
```
1    225
2     81
Name: surv_status, dtype: int64
```

In [3]:
```
#This seems to be an unbalanced data set
#for now treating this as balanced



#Lets analyse the 2D scatter plots  and see if something is evident

#since not much domain knowledge so lets check out possibilities
#About 75% of lymph from the breasts drains into the axillary lymph nodes, making

#Lets first take age vs operation_year

haberman.plot(kind="scatter",x='age',y='operation_year',label="age and operation_
plt.title("scatter plot-operation_year vs  age")
plt.legend()
plt.show()
```



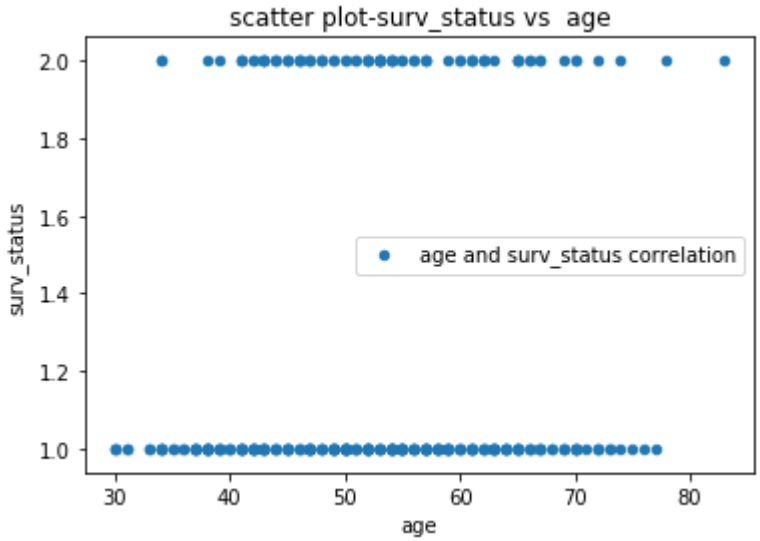

```
nothing much to comprehend; does not add value
```

In [4]:
```
#Lets check age vs axil_nodes
haberman.plot(kind="scatter",x='age',y='axil_nodes',label="age and axil_nodes cor
plt.title("scatter plot-axil_nodes vs  age")
plt.legend()
plt.show()
```



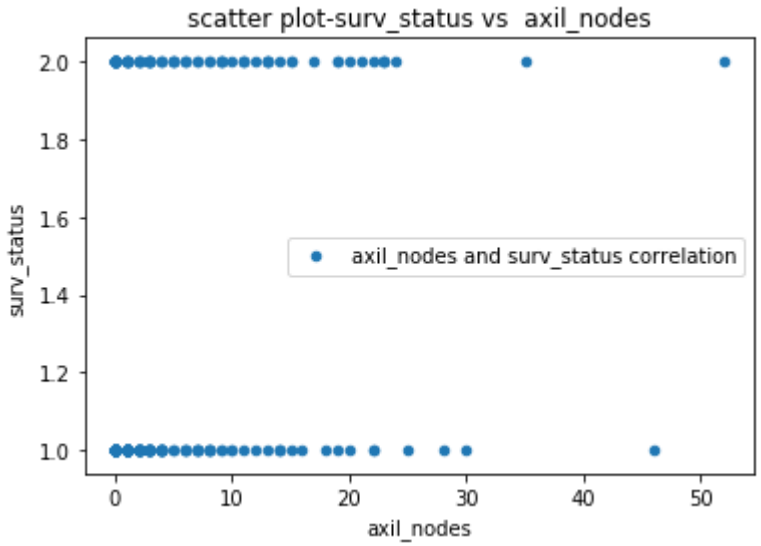

```
nothing very significant can be  concluded as such; the average axil_nodes is
less than 10 and very few have more than 30;; not very relevant to what we are
looking for
```

In [5]:
```
#Lets check age vs surv_status
haberman.plot(kind="scatter",x='age',y='surv_status',label="age and surv_status c
plt.title("scatter plot-surv_status vs  age")
plt.legend()
plt.show()
```



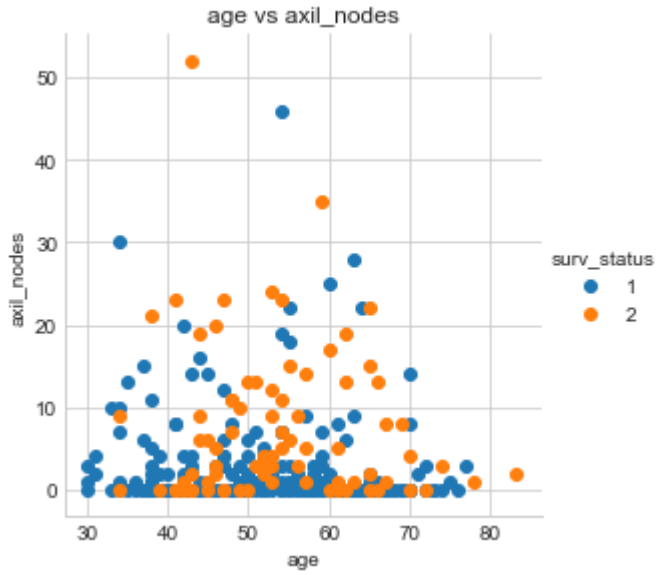

```
#not much to conclude
```

In [6]:
```
#lets check axil_nodes vs surv_status
haberman.plot(kind="scatter",x='axil_nodes',y='surv_status',label="axil_nodes and
plt.title("scatter plot-surv_status vs  axil_nodes")
plt.legend()
plt.show()
```



the number of axil_nodes usually varies between 0-30, but has no correlation with surv_status as such

In [7]:
```
# lets see a better view of age vs axil_nodes on surv_status
#color-code

sns.set_style("whitegrid")
sns.FacetGrid(haberman,hue="surv_status",size=4).map(plt.scatter,"age","axil_node
plt.title("age vs axil_nodes")
plt.show()
```



Observation:

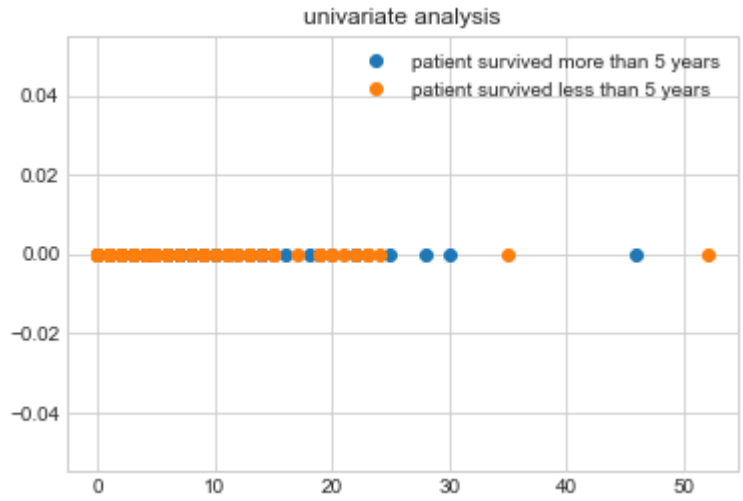
1. if the axil_nodes is greater than 50 then the patient dies before 5 years of operation (not much data to support)
2. very less patients have axil_nodes greater than 50
3. maximum patients have axil_nodes in the range of 0-20

In [18]:
```
#Lets see pair-plotting

plt.close()

sns.set_style("whitegrid")
sns.pairplot(haberman,hue="surv_status",size=3,vars=["age","operation_year","axil
plt.title("pair plots")

plt.show()
```

Observation: nothing much is evident from the pairplots, axil_nodes seems to be the best variable for it

In [9]:
```
#lets go for univariate analysis on axil_nodes

haberman_1=haberman.loc[haberman["surv_status"]==1]
haberman_2=haberman.loc[haberman["surv_status"]==2]
#print (haberman_1)
plt.plot(haberman_1["axil_nodes"],np.zeros_like(haberman_1["axil_nodes"]),'o',lab
plt.plot(haberman_2["axil_nodes"],np.zeros_like(haberman_2["axil_nodes"]),'o',lab
plt.legend()
plt.title("univariate analysis")
#plt.show()
```

Out[9]:
```
Text(0.5,1,'univariate analysis')
```

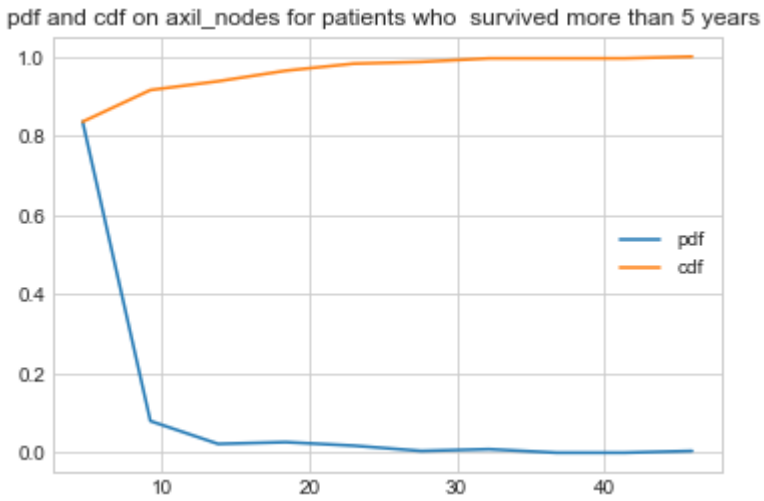


Observation: if axil nodes <= 15 then patients do not survive for more than 5 years after operation

```
sns.FacetGrid(haberman,hue="surv_status",size=5).map(sns.distplot,"axil_nodes")
.add_legend()
plt.show()
```

In [23]:
```
counts,bin_edges=np.histogram(haberman_1['axil_nodes'],bins=10,density=True)
pdf=counts/sum(counts)
print (pdf);
print (bin_edges)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label="pdf")
plt.plot(bin_edges[1:],cdf,label="cdf")
plt.legend()
plt.title("pdf and cdf on axil_nodes for patients who  survived more than 5 years
plt.show()
```

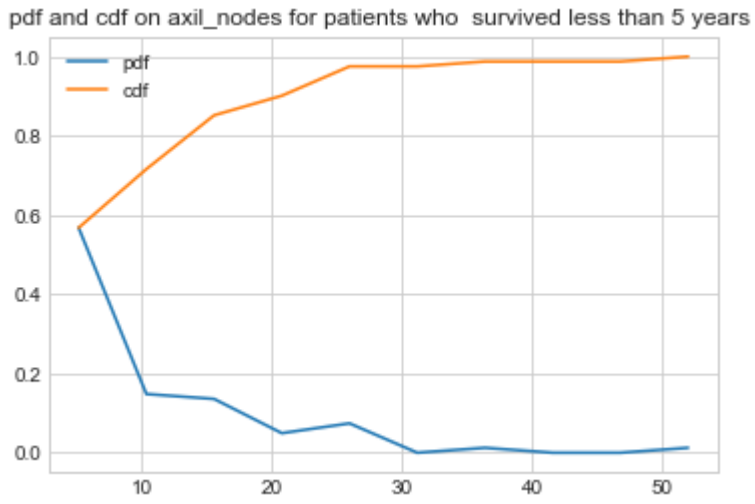
```
[0.83555556 0.08       0.02222222 0.02666667 0.01777778 0.00444444
 0.00888889 0.         0.         0.00444444]
[ 0.   4.6  9.2 13.8 18.4 23.  27.6 32.2 36.8 41.4 46. ]
```



people who survived more than 5 years do have less axil_nodes; majority have less than 10

In [11]:
```
counts,bin_edges=np.histogram(haberman_2['axil_nodes'],bins=10,density=True)
pdf=counts/sum(counts)
print (pdf);
print (bin_edges)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label="pdf")
plt.plot(bin_edges[1:],cdf,label="cdf")
plt.legend()
plt.title("pdf and cdf on axil_nodes for patients who  survived less than 5 years
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. ]
```



people who survived less than 5 years do have less axil_nodes too but they do have more axil_nodes on an average than people who survived more than 5 years

In [12]:
```
print ("Means:")
print (np.mean(haberman_1["axil_nodes"]))
print (np.mean(haberman_2["axil_nodes"]))

print ("Std-dev:")
print (np.std(haberman_1["axil_nodes"]))
print (np.std(haberman_2["axil_nodes"]))
```

```
Means:
2.7911111111111113
7.45679012345679
Std-dev:
5.857258449412131
9.128776076761632
```

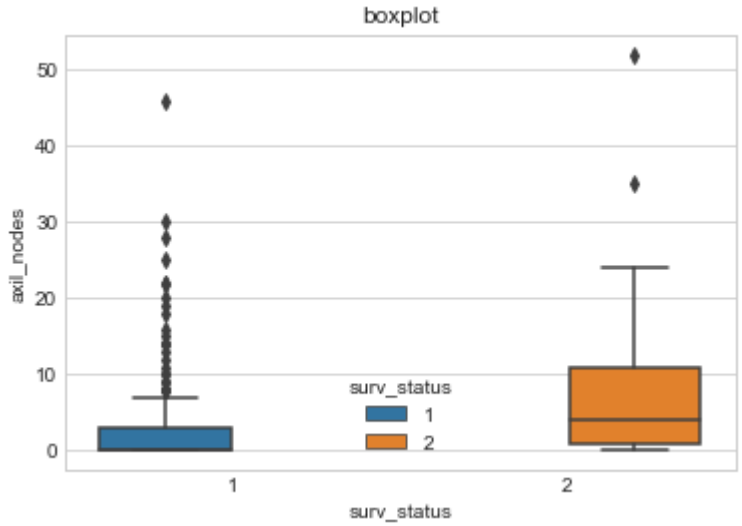
Observation: patients with axil_nodes around 2 live more than 5 years after operation

In [13]:
```
print ("Medians:")
print (np.median(haberman_1["axil_nodes"]))
print (np.median(haberman_2["axil_nodes"]))

print ("Quantiles:")
print (np.percentile(haberman_1["axil_nodes"],np.arange(0,100,25)))
print (np.percentile(haberman_2["axil_nodes"],np.arange(0,100,25)))

from statsmodels import robust
print ("Median Absolute Deviation:")
print (robust.mad(haberman_1["axil_nodes"]))
print (robust.mad(haberman_2["axil_nodes"]))
```

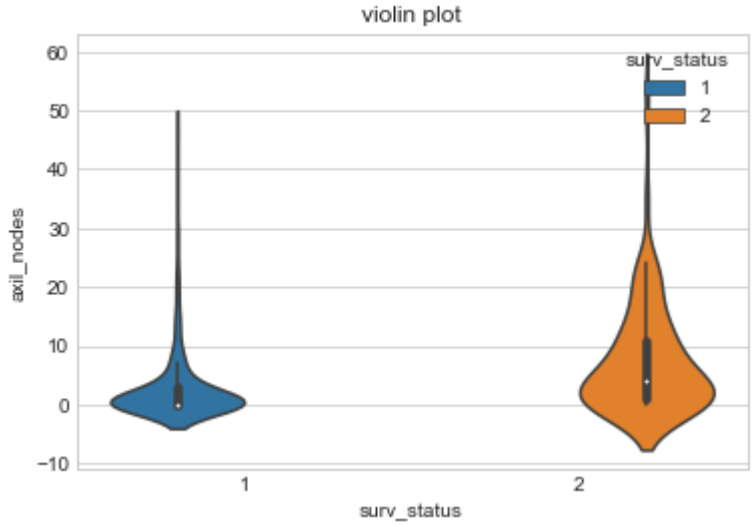
```
Medians:
0.0
4.0
Quantiles:
[0. 0. 0. 3.]
[ 0.  1.  4. 11.]
Median Absolute Deviation:
0.0
5.930408874022408
```

Type *Markdown* and LaTeX: α^2

In [21]:
```
#The medians and quantiles depict that there is not  much correlation with axil_n
# Lets see if box-plotting helps

sns.boxplot(x="surv_status",y="axil_nodes",data=haberman,hue="surv_status")
plt.title("boxplot")
#plt.legend()
plt.show()
```

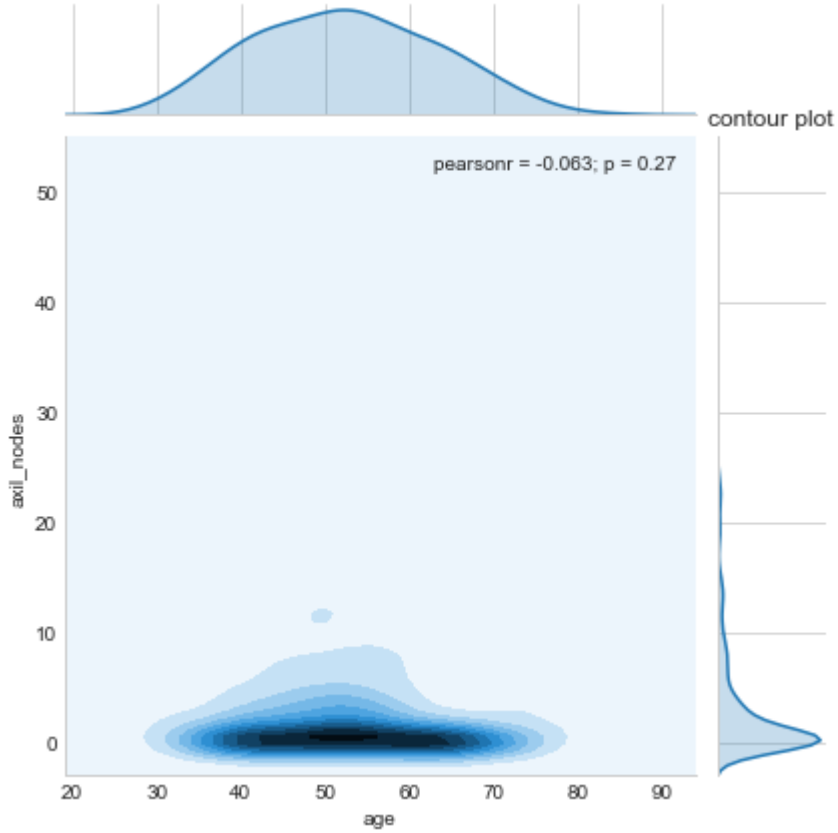


In [22]:
```
#Violin plots

sns.violinplot(x="surv_status",y="axil_nodes",data=haberman,size=4,hue="surv_stat
plt.title("violin plot")
plt.show()
```



Joint plot

In [16]:
```
sns.jointplot(x="age",y="axil_nodes",data=haberman,kind="kde")
plt.title("contour plot")
plt.show()
```



axil_nodes may not be a deciding factor for survey_status

I started with getting the right deciding factor to predict the survey_status on the basis of the age,operation_year and axil_nodes. Started with scatter plot to see if there was some correlation between age and operation_year or axil_nodes. Could not zero it down to something significant. The scatter plot for axil_nodes vs age led to the following - Observation:

1. if the axil_nodes is greater than 50 then the patient dies before 5 years of operatio
2. very less patients have axil_nodes greater than 50
3. maximum patients have axil_nodes in the range of 0-20

Pair-plotting also does not give anything significant. Some correlation was found between axil_nodes and survival_status : Observation: if axil nodes <= 15 then patients do not survive for more than 5 years after operation

Observing the histogram and the pdf's , it seems like patients with axil_nodes around 2 live more than 5 years after operation.

Box-plot and violing plots could only lead to the conclusion -

axil_nodes may not be a deciding factor for survey_status