In [1]:

```
import pandas as pd
import numpy as np
import os
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")


#importing the relevant packages
#Load haberman.csv into a pandas dataFrame.
haber=pd.read_csv("haberman.csv")
```

ANALYSIS OF DATA SHAPE AND FEATURES

In [2]:

```
print(haber.shape)
#shape of the file..number of data points and features
```

```
(306, 4)
```

In [3]:

```
print(haber.columns)
```

```
Index(['age', 'op_year', 'axil_nodes', 'surv_status'], dtype='object')
```

In [4]:

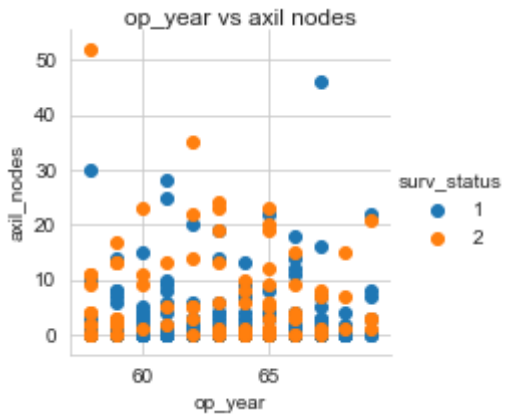
```
#no of data points presentin the column present in the data set
haber["surv_status"].value_counts()
#thus unbalaced data set
```

Out[4]:

```
1    225
2     81
Name: surv_status, dtype: int64
```

In [6]:

```
sns.set_style("whitegrid")
g=sns.FacetGrid(haber,hue="surv_status",size=3) \
   .map(plt.scatter,"op_year","axil_nodes").add_legend()
g.fig.suptitle('op_year vs axil nodes')
plt.show(g)
#PEOPLE WHO  DONT SURVIVE HAVE A SLIGHTLY HIGH NUMBER OF AXIL_NODES THAN THSE WHO SURVI
VE
```



ANALYSING PAIRPLOTS TO GET A BETTER PICTURE OF DEPENDENCIES

In [9]:

```
# let us try pair plot
plt.close()
g=sns.set_style("whitegrid")
g=sns.pairplot(haber,hue="surv_status",vars=['age','axil_nodes','op_year'],size=3)
g.fig.suptitle('PAIRPLOTS')
plt.show(g);
```

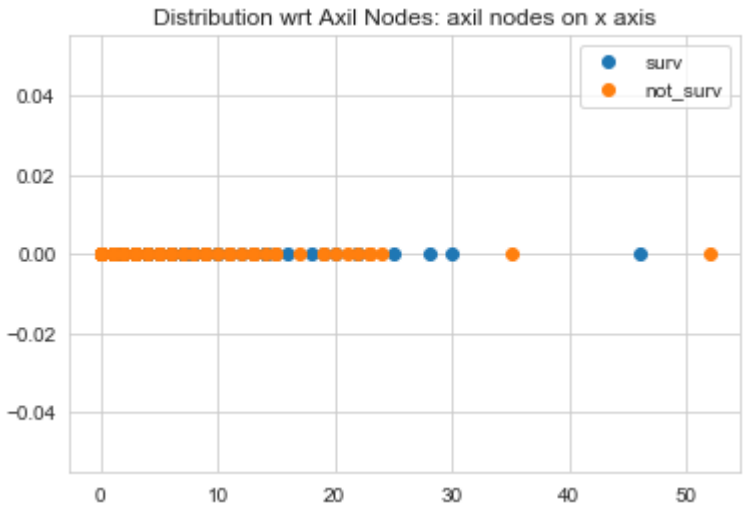
CONCUSIONS FROM PAIR PLOTS/:

PEOPLE WITH LESS AGE ARE MORE LIKELY TO SURVIVE PEOPLE WITH AN EARLY OPERATION ARE HAVING A DLIGHT ADVANTAGE IN CHANCES OF SURVIVAL PEOPLE WITH MORE AXIL NODES ARE LESS LIKELY TO SURVIVE

ANALYSING DATA DISTRIBUTION

In [14]:

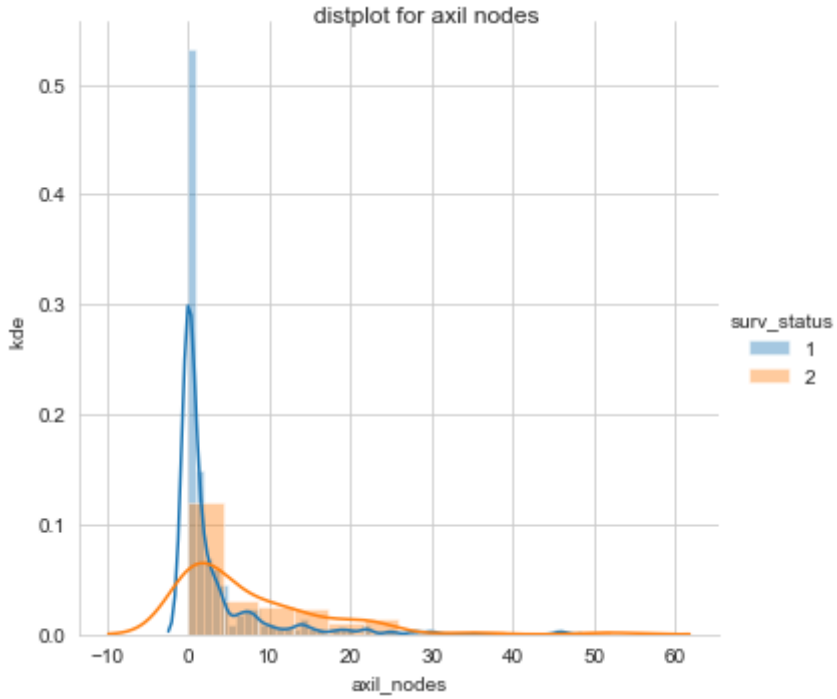
```
#DATA DISTRIBUTION WITH RESPECCT TO AXIL  NODES
import numpy as np

survive=haber.loc[haber["surv_status"]==1]
not_survive=haber.loc[haber["surv_status"]==2]
plt.plot(survive["axil_nodes"],np.zeros_like(survive["axil_nodes"]),'o',label="surv")
plt.plot(not_survive["axil_nodes"],np.zeros_like(not_survive["axil_nodes"]),'o',label=
"not_surv")
plt.ylabel=('axil_nodes')
plt.xlabel=('none')
plt.title("Distribution wrt Axil Nodes: axil nodes on x axis")
plt.legend()
plt.show()
```



WE CAN CLEARLY SEE THAT AS THE AGE INCREASES THE SURVIVAL CHANCSE RE SOMEWHAT LESS BUT STILL THERE ARE CONSIDERABLE NUMBER OF POINTS TO CONSIDER A MEDICAL DIAGNOSIS...

In [16]:

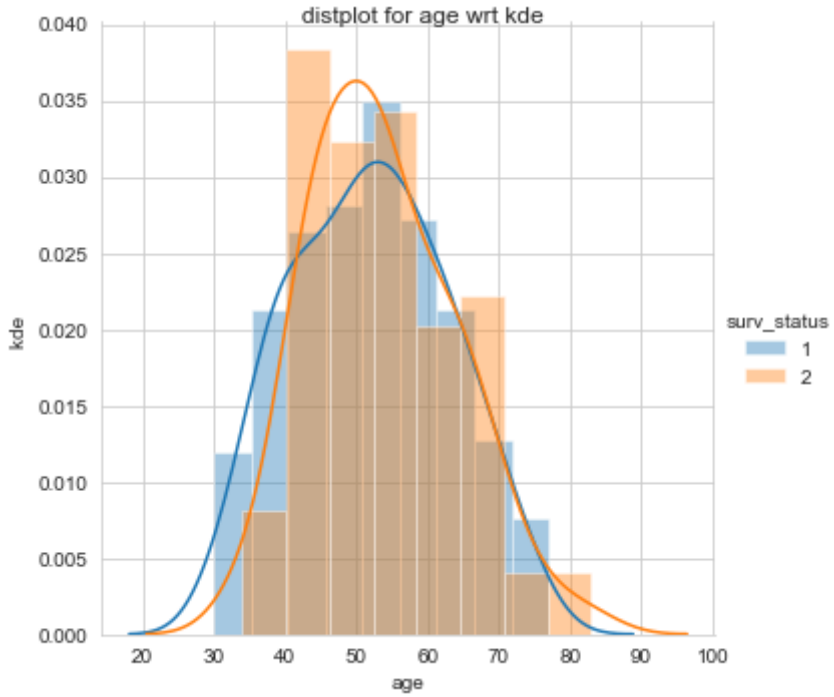
```
#KDE PLOTS

g=sns.FacetGrid(haber,hue="surv_status",size=5) \
    .map(sns.distplot,"axil_nodes").add_legend()
g.set(xlabel = 'axil_nodes', ylabel = 'kde ')
g.fig.suptitle('distplot for axil nodes')
plt.show(g)
#when there are no axil nodes ,then the survIval chances are pretty high,frm the blue l
ine
#as the no of axil nodes are above 10 tthe surviva; chances are drastically reduced....
```



WE CAN OBSERVE THAT THE AXIL NODES ARE USUALLY HIGHER IN PEOPLE WHO DONT SURVIVE.ALSO THE PEOPLE WHO SURVIVE USUALLLY GET RID OF NDES AFTER A SHORT PERIOD OF TIME

In [17]:

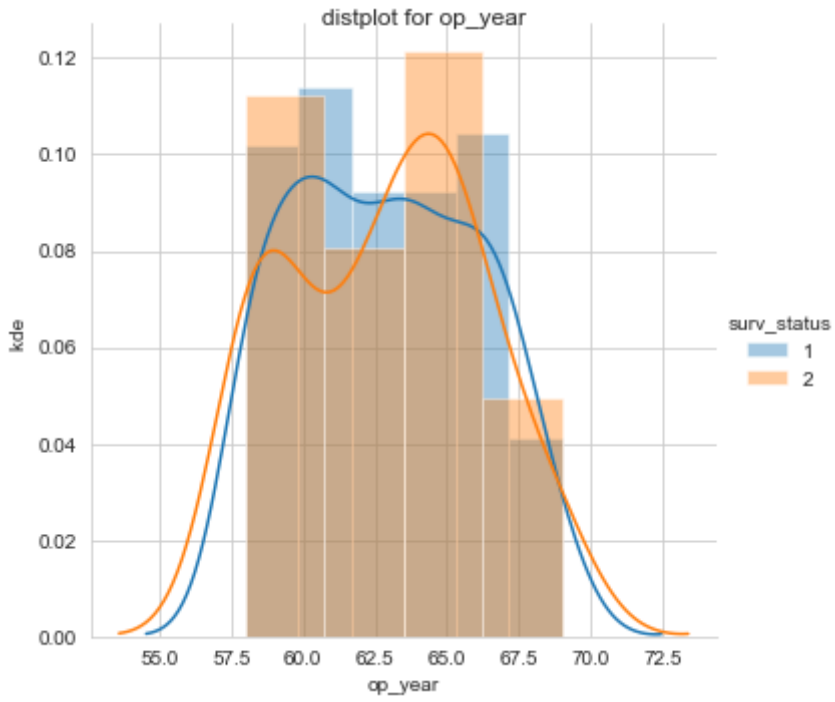
```
#KDE PLOT

g=sns.FacetGrid(haber,hue="surv_status",size=5) \
    .map(sns.distplot,"age").add_legend()
g.set(xlabel = 'age', ylabel = 'kde ')
g.fig.suptitle('distplot for age wrt kde')
plt.show()

#Lesser aged people have sliightly more chance of survival
```



THIS PLOT SHOWS US THE AGE DISTRIBUTION OF THE TWO CATEGORIES IN SURVIVAL.

In [18]:

```
g=sns.FacetGrid(haber,hue="surv_status",size=5) \
    .map(sns.distplot,"op_year").add_legend()
g.set(xlabel = 'op_year', ylabel = 'kde ')
g.fig.suptitle('distplot for op_year')
plt.show()

#peoplewho get operation early are more likely to survive
```

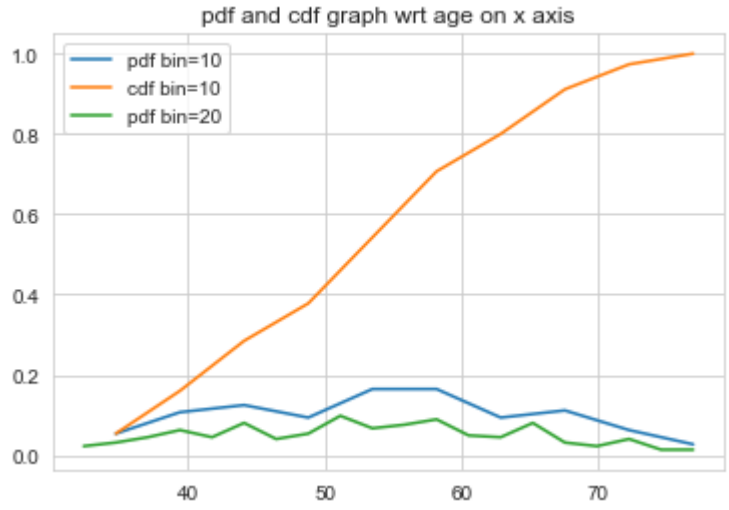


THIS PLOT SHOWS US THE DISTRIBUTION IN WHICH BOTH TYPE OF PATIENTS GOT OPERATED

In [29]:

```
#PLOTTING PDF AND CDF

import numpy as np
counts,binedges=np.histogram(survive['age'],bins=10,density=True)
pdf=counts/(sum(counts))
print(counts)
print(pdf)
print(binedges)
cdf=np.cumsum(pdf)
plt.plot(binedges[1:],pdf,label='pdf bin=10')
plt.plot(binedges[1:],cdf,label='cdf bin=10')

counts,binedges=np.histogram(survive['age'],bins=20,density=True)
pdf=counts/(sum(counts))
g=plt.plot(binedges[1:],pdf,label='pdf bin=20')
plt.legend()
plt.title('pdf and cdf graph wrt age on x axis')
plt.show()
#pdf distribution of people w.r.t age
#the data is prettymuch linear so the disease affects people of different age equally
```

```
[0.01134752 0.02269504 0.02647754 0.01985816 0.03498818 0.03498818
 0.01985816 0.02364066 0.01323877 0.00567376]
[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. ]
```



PDF AND CDF PLOT OF DIFFERENT AGE GROUP OF PEOPLE the data is prettymuch linear so the disease affects people of different age equally

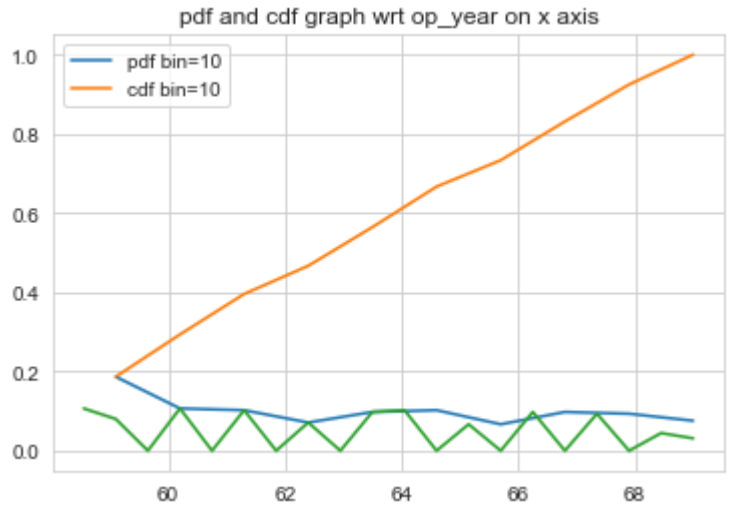
In [31]:

```
#PLOTTING PDF NADD CDF FOR OP_YEAR

import numpy as np
counts,binedges=np.histogram(survive['op_year'],bins=10,density=True)
pdf=counts/(sum(counts))
print(counts)
print(pdf)
print(binedges)
cdf=np.cumsum(pdf)
plt.plot(binedges[1:],pdf,label='pdf bin=10')
plt.plot(binedges[1:],cdf,label='cdf bin=10')

counts,binedges=np.histogram(survive['op_year'],bins=20,density=True)
pdf=counts/(sum(counts))
plt.legend()
plt.title('pdf and cdf graph wrt op_year on x axis')
plt.plot(binedges[1:],pdf,label='pdf bin=20')
```

```
[0.16969697 0.0969697  0.09292929 0.06464646 0.08888889 0.09292929
 0.06060606 0.08888889 0.08484848 0.06868687]
[0.18666667 0.10666667 0.10222222 0.07111111 0.09777778 0.10222222
 0.06666667 0.09777778 0.09333333 0.07555556]
[58.  59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
```

Out[31]:

```
[<matplotlib.lines.Line2D at 0x18a1701e320>]
```



PDF AND CDF PLOT OF OPERATION YEAR

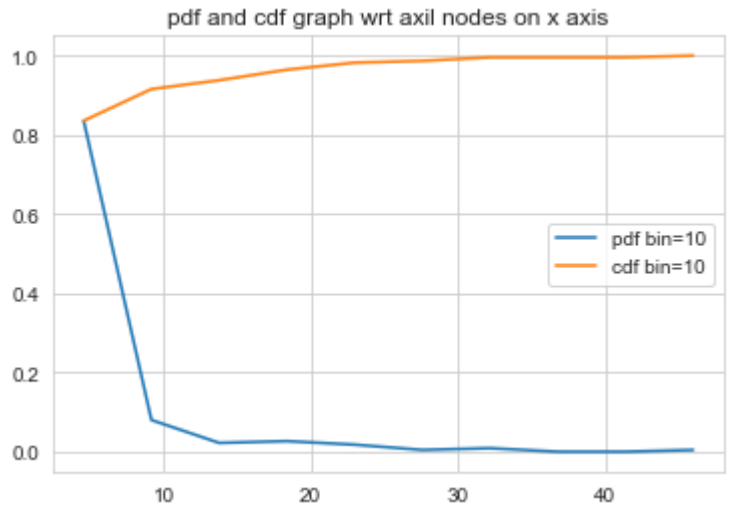
In [36]:

```
#PLOTTING PDF NADD CDF FOR AXIL_NODES


import numpy as np
counts,binedges=np.histogram(survive['axil_nodes'],bins=10,density=True)
pdf=counts/(sum(counts))
print(counts)
print(pdf)
print(binedges)
cdf=np.cumsum(pdf)

plt.plot(binedges[1:],pdf,label='pdf bin=10')
plt.plot(binedges[1:],cdf,label='cdf bin=10')
plt.legend()
plt.title('pdf and cdf graph wrt axil nodes on x axis')
```

```
[0.18164251 0.0173913  0.00483092 0.0057971  0.00386473 0.00096618
 0.00193237 0.         0.         0.00096618]
[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. ]
```

Out[36]:

```
Text(0.5, 1.0, 'pdf and cdf graph wrt axil nodes on x axis')
```



PDF AND CDF PLOT OF AXIL NODES

In [40]:

```
#Mean, Variance, Std-deviation,
print("Means:")
print(np.mean(survive["axil_nodes"]))
#Mean with an outlier.
print(np.mean(np.append(survive["axil_nodes"],50)));
print(np.mean(not_survive["axil_nodes"]))

print("\nStd-dev:");
print(np.std(survive["axil_nodes"]))
print(np.std(not_survive["axil_nodes"]))
```

```
Means:
2.7911111111111113
3.0
7.45679012345679

Std-dev:
5.857258449412131
9.128776076761632
```

AS WE CAN SEE THE MEAN NUMBER OF NODES FOR SURVIVING PEOPLE IS LESS THAN THAT OF NON-SURVING PEOPLE

In [41]:

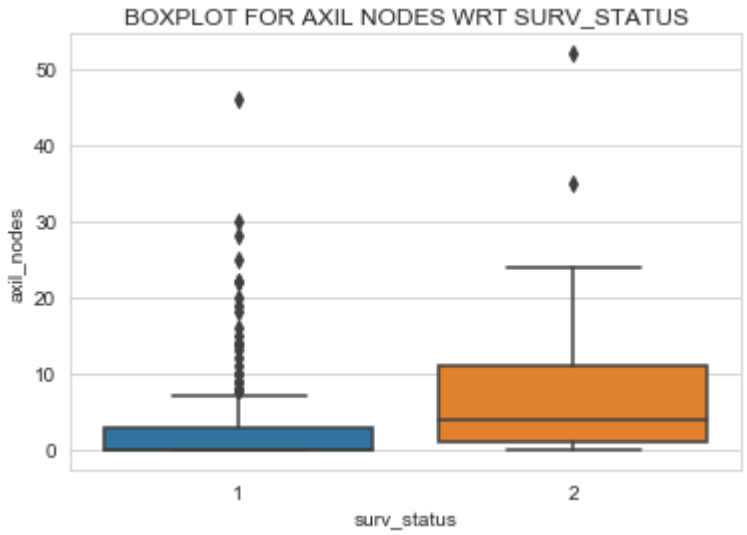
```
#Median, Quantiles, Percentiles, IQR.
print("\nMedians:")
print(np.median(survive["axil_nodes"]))
#Median with an outlier
print(np.median(np.append(survive["axil_nodes"],50)));
print(np.median(not_survive["axil_nodes"]))


print("\nQuantiles:")
print(np.percentile(survive["axil_nodes"],np.arange(0, 100, 25)))
print(np.percentile(not_survive["axil_nodes"],np.arange(0, 100, 25)))

print("\n90th Percentiles:")
print(np.percentile(survive["axil_nodes"],90))
print(np.percentile(not_survive["axil_nodes"],90))

#gives the mean absolute deviation along the speciied axis
from statsmodels import robust
print ("\nMedian Absolute Deviation")
print(robust.mad(survive["axil_nodes"]))
print(robust.mad(not_survive["axil_nodes"]))
```

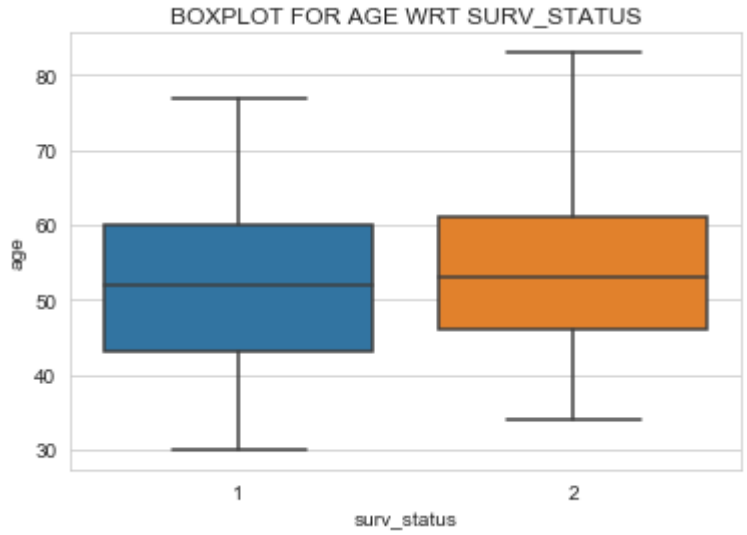
```
Medians:
0.0
0.0
4.0

Quantiles:
[0. 0. 0. 3.]
[ 0.  1.  4. 11.]

90th Percentiles:
8.0
20.0

Median Absolute Deviation
0.0
5.930408874022408
```

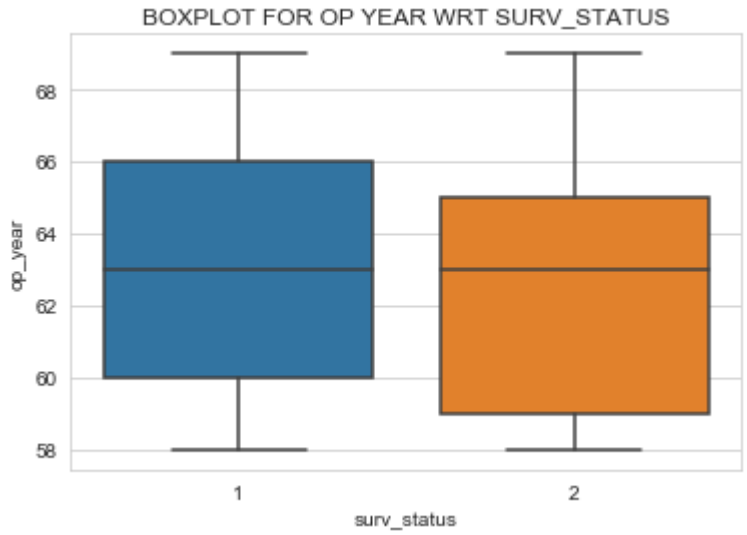
MEAN,MEDIAN AND ABSOLUTE DEVIATION

In [48]:

```
# IN the plot below, a technique call inter-quartile range is used in plotting the whis
kers.
#Whiskers in the plot below donot correposnd to the min and max values.

#Box-plot can be visualized as a PDF on the side-ways.

g=sns.boxplot(x='surv_status',y='axil_nodes', data=haber)
#no need for legend
plt.title('BOXPLOT FOR AXIL NODES WRT SURV_STATUS')
plt.show(g)
sns.boxplot(x='surv_status',y='age', data=haber)
plt.title('BOXPLOT FOR AGE WRT SURV_STATUS')

plt.show()
sns.boxplot(x='surv_status',y='op_year', data=haber)
plt.title('BOXPLOT FOR OP YEAR WRT SURV_STATUS')

plt.show()
```



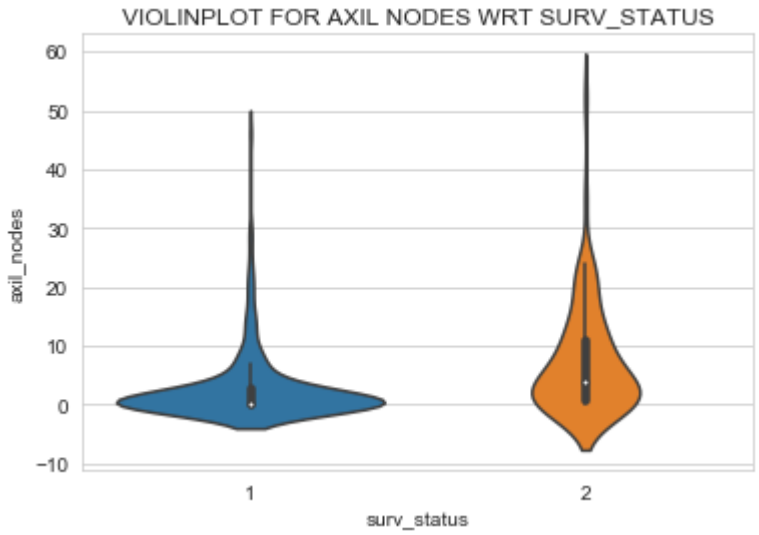


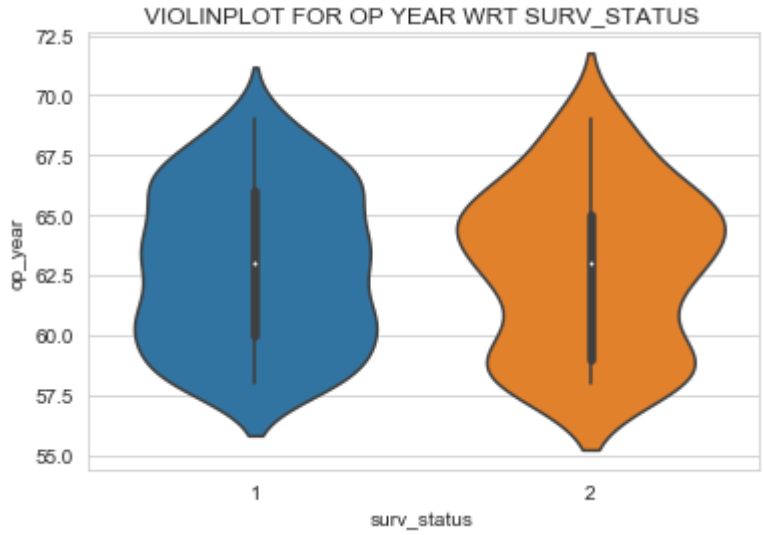


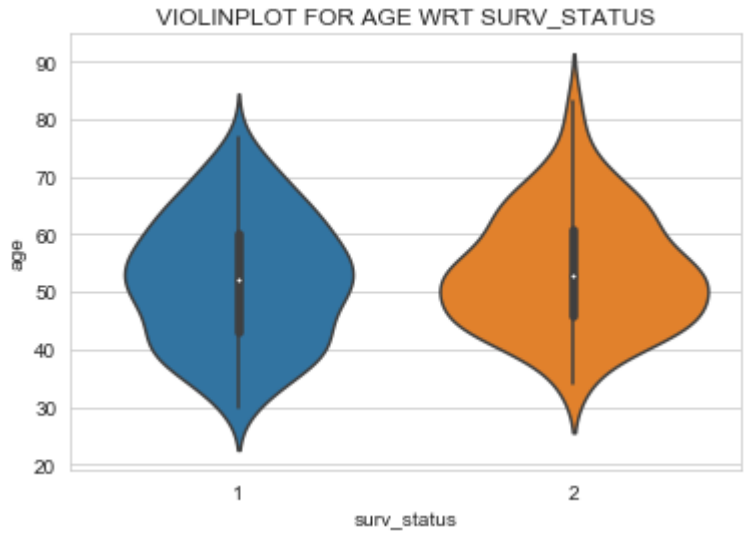
BOXPLOT DIAGRAMS FOR VARIOUS FEATURES AND THEIR DISTRIBUTION

In [49]:

```
###Violin plots####

sns.violinplot(x="surv_status", y="axil_nodes", data=haber, size=10)
plt.title('VIOLINPLOT FOR AXIL NODES WRT SURV_STATUS')

plt.show()
sns.violinplot(x="surv_status", y="op_year", data=haber, size=10)
plt.title('VIOLINPLOT FOR OP YEAR WRT SURV_STATUS')

plt.show()
sns.violinplot(x="surv_status", y="age", data=haber, size=10)
plt.title('VIOLINPLOT FOR AGE WRT SURV_STATUS')

plt.show()
```



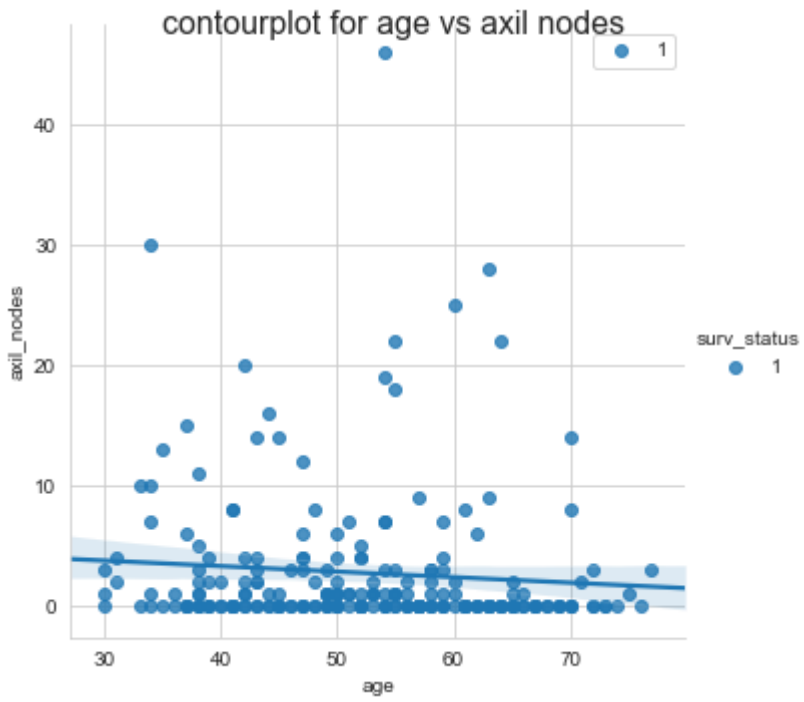




VIOLIN PLOTS FOR DATA DISTRIBUTION

In [62]:

```
#2D Density plot, contors-plot
g=sns.lmplot(x="age", y="axil_nodes",hue='surv_status', data=survive,fit_reg='False');
plt.legend()
plt.suptitle('contourplot for age vs axil nodes', fontsize = 16)
plt.show(g);
#2D Density plot, contors-plot
#sns.jointplot(x="age", y="axil_nodes", data=not_survive, kind="reg");
#plt.show();
```



CLEARLY WITH AGE THE AXIL NODES ARE INCREASING .THAT IS OUR TAKEAWAY FROM THIS PLOT

CONCLUSION

A conclusion which can be derived is that the survivial status depends somewhat more n the number of axil nodes but overall the data is very overlapping and no concreate observation can be made

it can be said that people with more age and axil nodes are more likely not to survive an early operation can give people some hope of surviving maybe the data can be processed to give us a better picture