In [2]:

```
#Here we have Haberman's survival data set
#according to given dataset..we mainly have four attributes which are as follows
#1)Age of patient at time of operation (numerical)
#2)Patient's year of operation (year - 1900, numerical)
#3)Number of positive axillary nodes detected (numerical)
#4)Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the
 patient died within 5 year

# here no. of instances are 306

# Axillary node:The axillary nodes are a group of lymph nodes located in the axillary
 (or armpit) region of the body.
#They perform the vital function of filtration and conduction of lymph from the upper l
imbs, pectoral region, and upper back.


import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
#Load haberman.csv into a pandas dataFrame.
Haberman = pd.read_csv("haberman.csv")
```

In [5]:

```
# (Q) how many data-points and featrues are there?
print (Haberman.shape)
```

```
(306, 4)
```

In [6]:

```
#(Q) What are the column names in our dataset?
print (Haberman.columns)
```

```
Index([u'age', u'year', u'axillary', u'survived'], dtype='object')
```

In [8]:

```
#(Q) How many data points for each class are present?

Haberman["survived"].value_counts()
```

Out[8]:

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

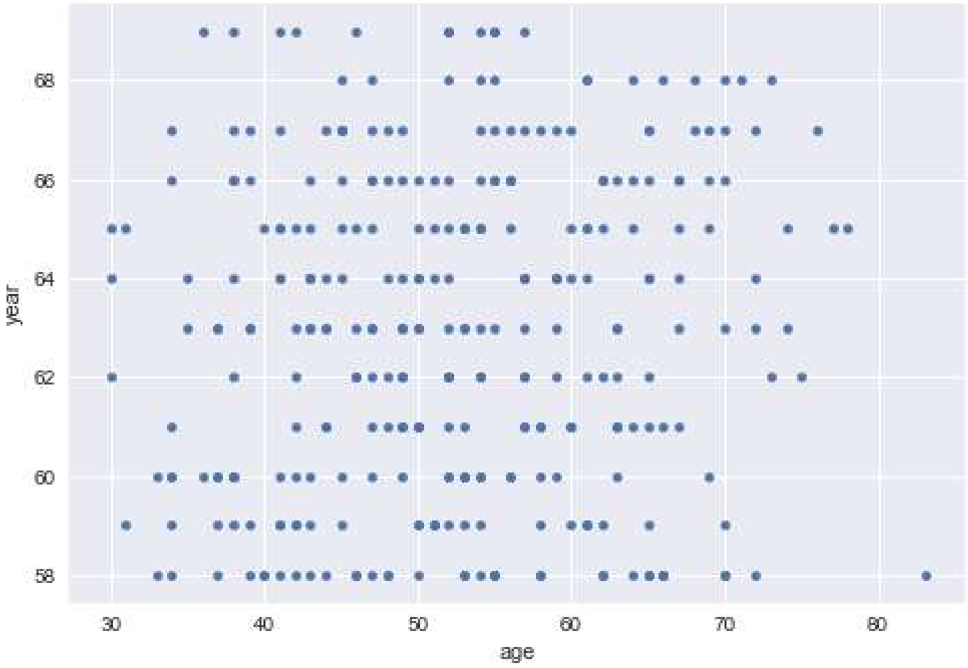
In [ ]:

```
#By seeing the above datapoints for each classes we comes on some conclusion that
# 225 patients are survived 5 or greater than 5 year,
# and 81 are died within 5 year of there operation

#Now we have four attributes..so lets draw some graph to find some information from tha
t.
```

2D SCATTER PLOT

In [10]:

```
#2D SCATTER PLOT

Haberman.plot(kind='scatter', x='age', y='year') ;
plt.show()
```

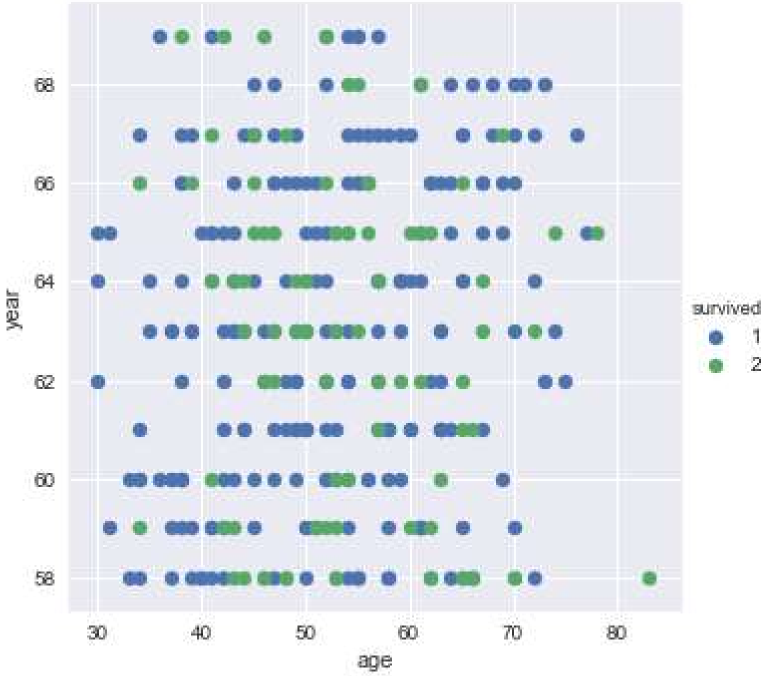


In [ ]:

```
#cannot make any useful conclusion by seeing lets try some color coding then see it
```

In [11]:

```
# Here 'sns' corresponds to seaborn.
sns.FacetGrid(Haberman, hue="survived", size=5) \
   .map(plt.scatter, "age", "year") \
   .add_legend();
plt.show();
```

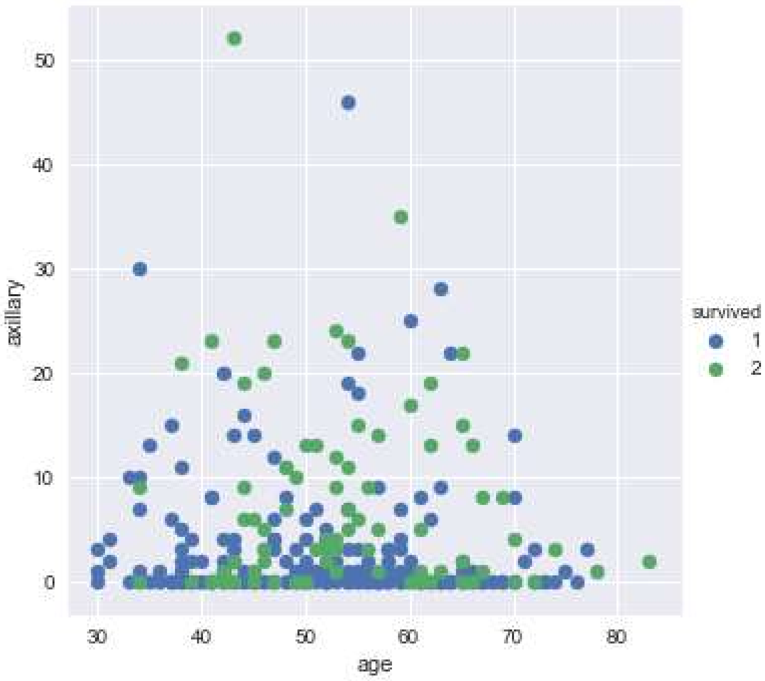


In [ ]:

```
#its not easily separable..so try some other attributes
```

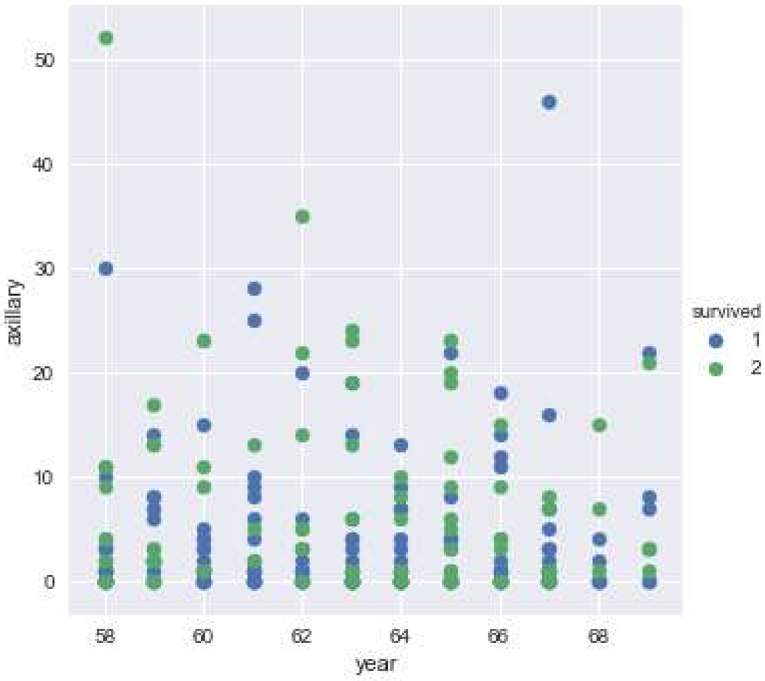
In [12]:

```
# Here 'sns' corresponds to seaborn.
sns.FacetGrid(Haberman, hue="survived", size=5) \
   .map(plt.scatter, "age", "axillary") \
   .add_legend();
plt.show();
```



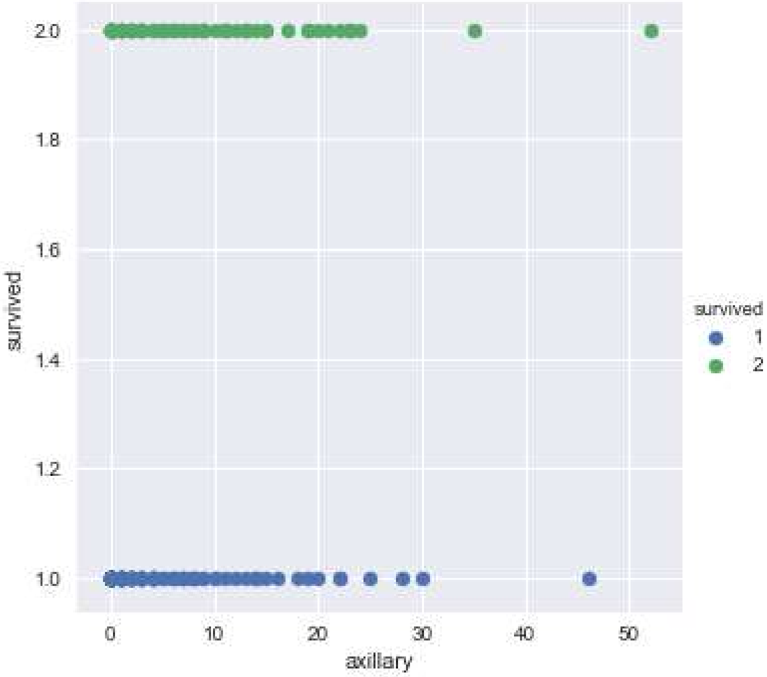
In [13]:

```
# Here 'sns' corresponds to seaborn.
sns.FacetGrid(Haberman, hue="survived", size=5) \
   .map(plt.scatter, "year", "axillary") \
   .add_legend();
plt.show();
```



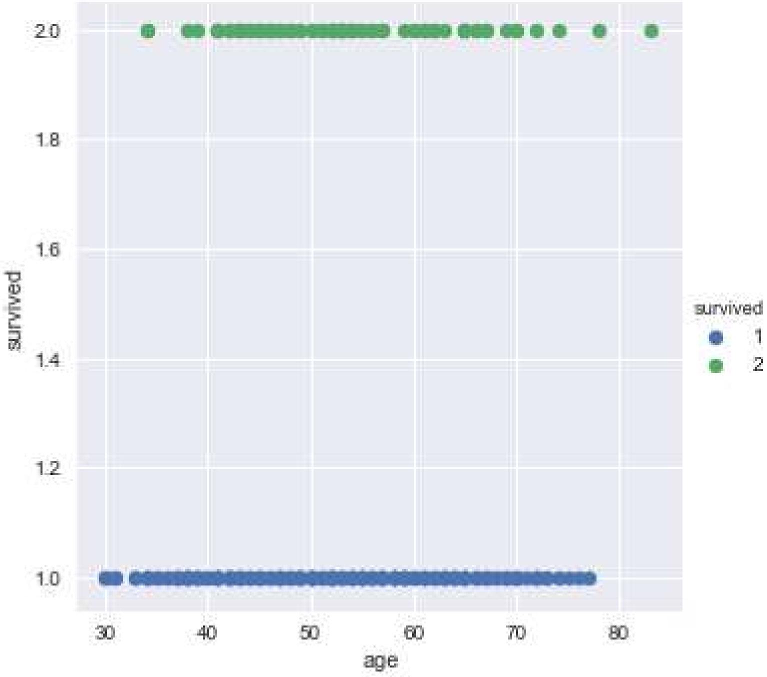
In [14]:

```
# Here 'sns' corresponds to seaborn.
sns.FacetGrid(Haberman, hue="survived", size=5) \
   .map(plt.scatter, "axillary", "survived") \
   .add_legend();
plt.show();
```



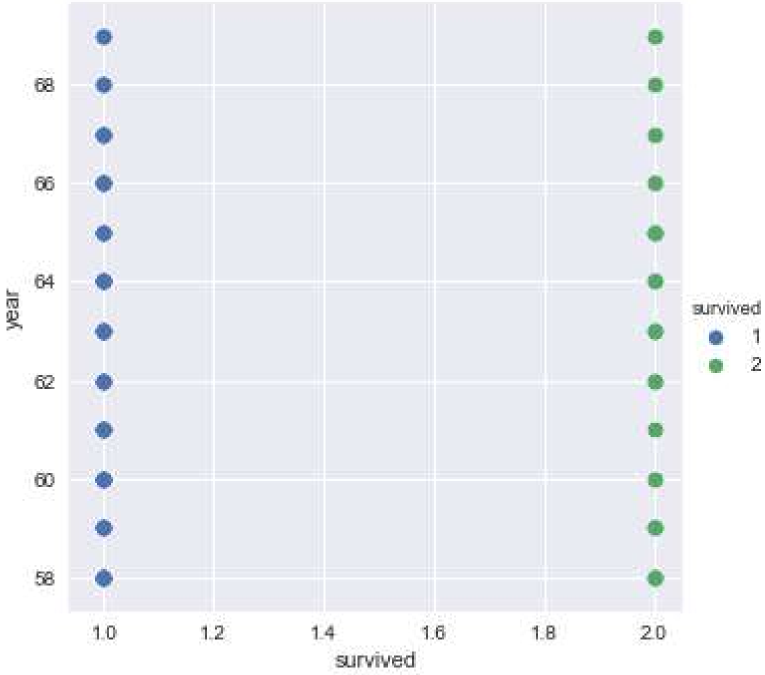
In [15]:

```
# Here 'sns' corresponds to seaborn.
sns.FacetGrid(Haberman, hue="survived", size=5) \
   .map(plt.scatter, "age", "survived") \
   .add_legend();
plt.show();
```



In [16]:

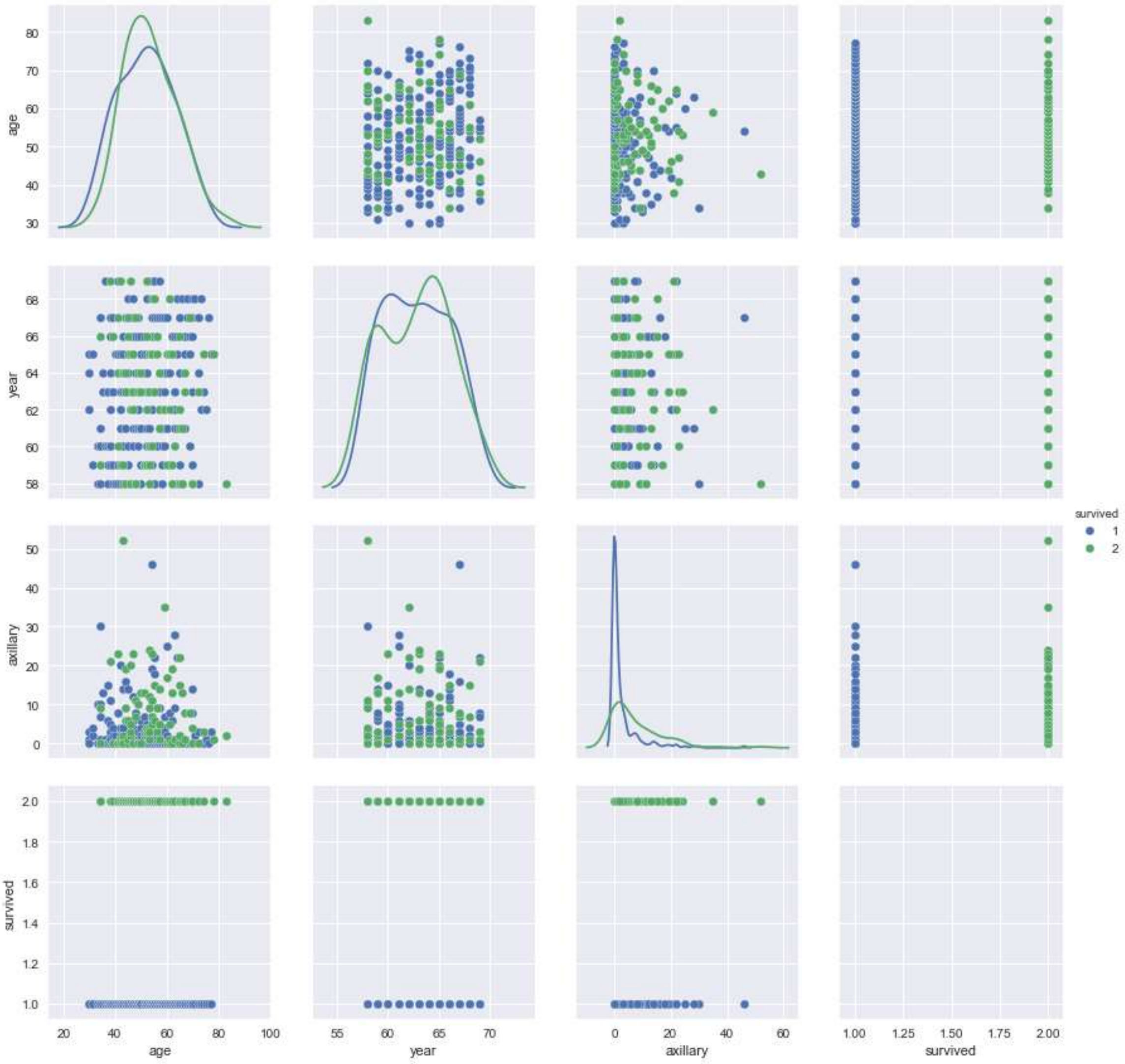
```
# Here 'sns' corresponds to seaborn.
sns.FacetGrid(Haberman, hue="survived", size=5) \
   .map(plt.scatter, "survived", "year") \
   .add_legend();
plt.show();
```



Pair-plot

In [29]:

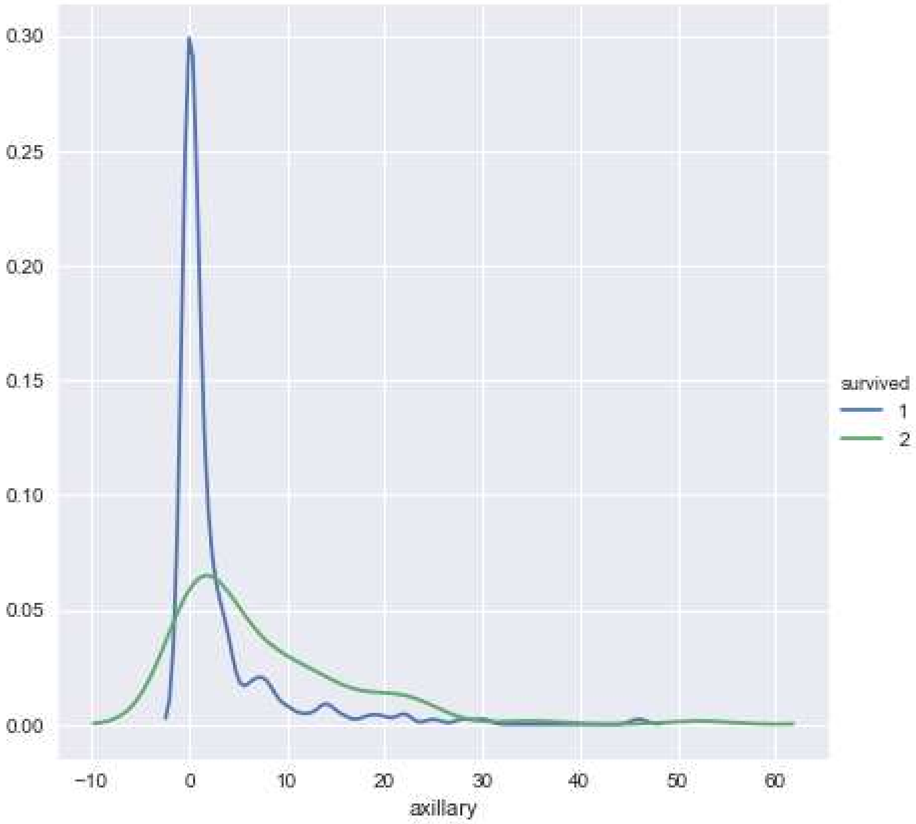
```
# pairwise scatter plot: Pair-Plot
# Dis-advantages:
##Can be used when number of features are high.
##Cannot visualize higher dimensional patterns in 3-D and 4-D. Only possible to view 2D
 patterns.
plt.close();
sns.pairplot(Haberman,hue="survived", size=3, diag_kind="kde");
plt.show()


# NOTE: the diagnol elements are PDFs for each feature. PDFs are expalined below.
```



PDF

In [30]:

```
#Seaborn plot of petal_length's PDF.
sns.FacetGrid(Haberman,hue="survived", size=6) \
   .map(sns.kdeplot, "axillary") \
   .add_legend();
plt.show();
```



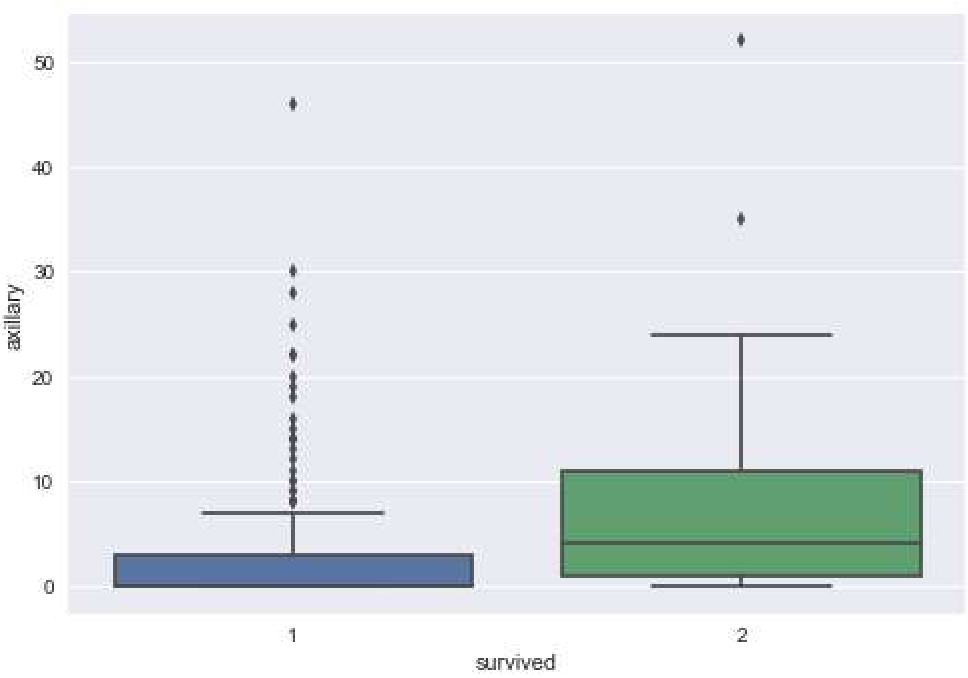
In [ ]:

```
#by seeing above pdf we have only one conclusion that "if no. of axillary node are less
 than chance of survival is more"
```

Box-Plot

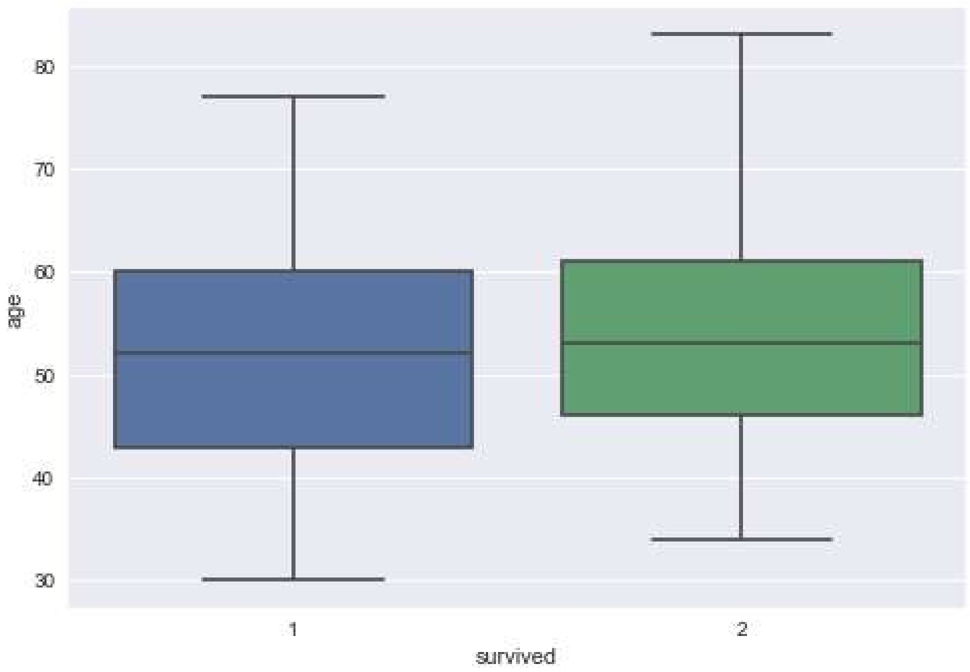
In [31]:

```
# Box-plot with whiskers: another method of visualizing the  1-D scatter plot more intuitivey.
# The Concept of median, percentile, quantile.
# How to draw the box in the box-plot?
# How to draw whiskers: no standard way. Could use min and max or use other complex statistical techniques.

#NOTE: IN the plot below, a technique call inter-quartile range is used in plotting the whiskers.
#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.

sns.boxplot(x='survived',y='axillary', data=Haberman)
plt.show()
```



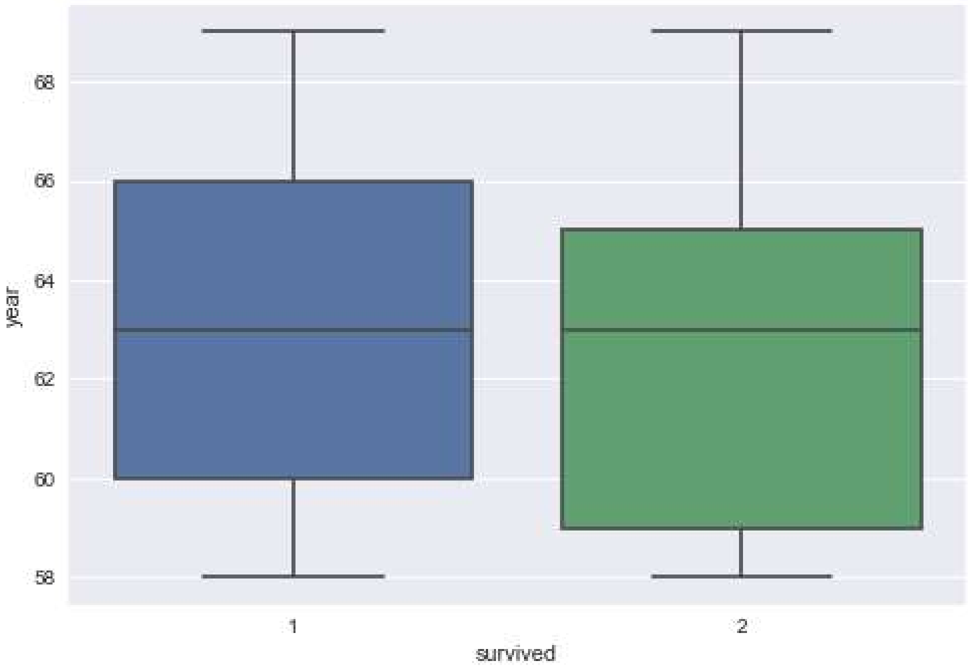
In [32]:

```
sns.boxplot(x='survived',y='age', data=Haberman)
plt.show()
```



In [33]:

```
sns.boxplot(x='survived',y='year', data=Haberman)
plt.show()
```



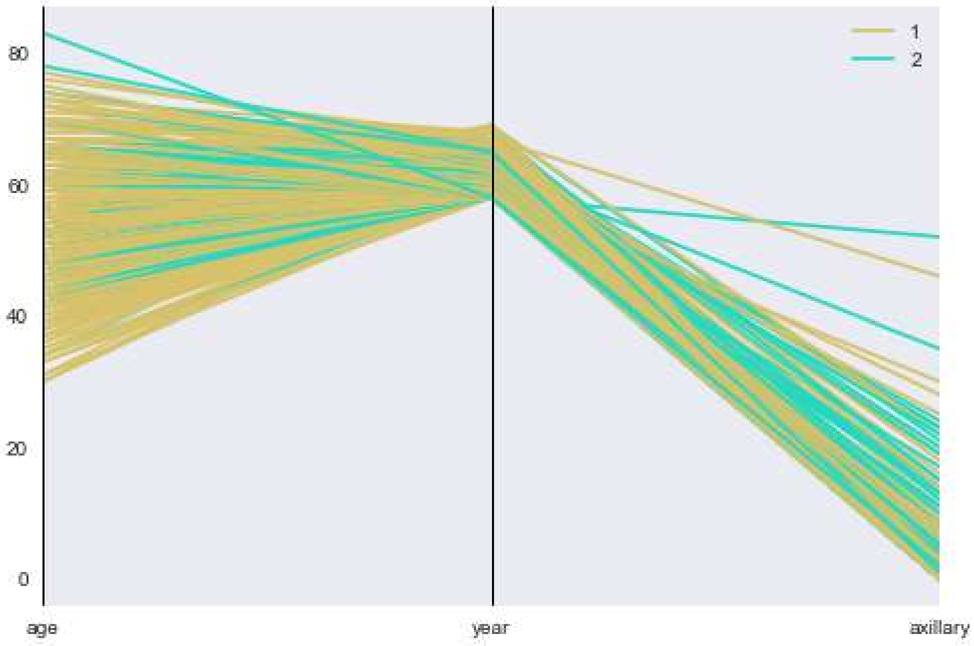
In [ ]:

```
#by seeing above box-plots we didn't comes on any conclusion,it has mixed upto 50 to 70
 percent
```

PARALLEL COORDINATES PLOT

In [35]:

```
# Parallel-coordinates to visualizae data when we have more than 5 dimensions and pair-plot is too ahrd to understand.
# How to draw a parallel-coordinate plot.
# Diasadv: crowding, whats the correct order of the features/dimensions.

from pandas.plotting import parallel_coordinates
parallel_coordinates(Haberman, "survived");
plt.show();
```



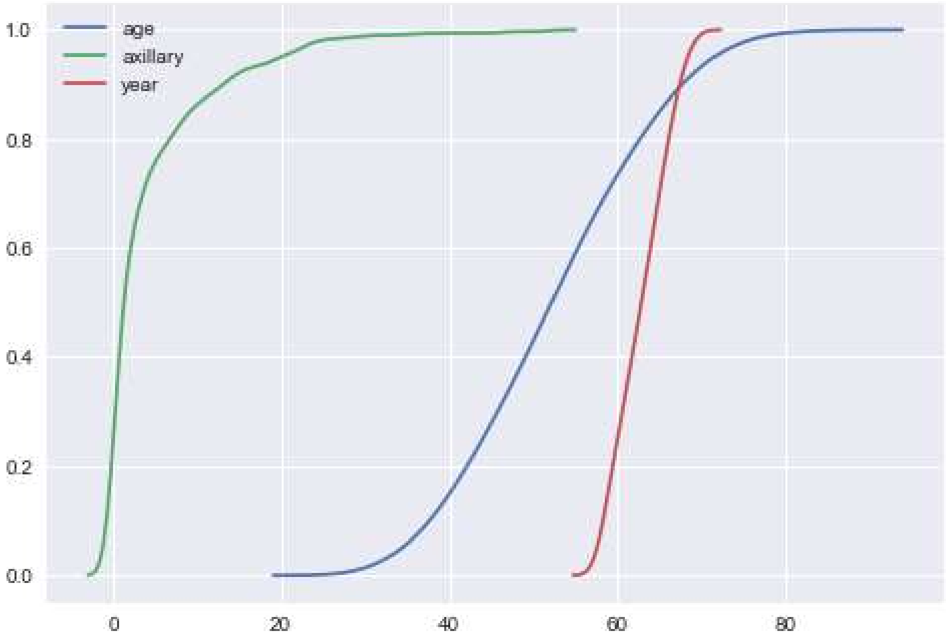
In [ ]:

```
#here also we didn't comes on any conclusion
```

CDF

In [36]:

```
ax = sns.kdeplot(Haberman['age'], cumulative=True)
ax = sns.kdeplot(Haberman['axillary'], cumulative=True)
ax = sns.kdeplot(Haberman['year'], cumulative=True)
plt.show()
```



Conclusion

In [ ]:

```
# By plotting all pdf,cdf,box-plot,pair plots,scatter plot etc. we get only one conclus
ion :
#"if no. of axillary node is less,than survival of patients is more"

#We need more features to comes on very good conlusion
```