In [4]:
```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
Filelocation="C:\\Users\\A674682\\Desktop\\work&not\\ipynbfiles\\haberman.csv"
Filelocation1="C:\\Users\\akars_wj3kiua\\Downloads\\haberman.csv"
```

In [5]:
```python
Hdata=pd.read_csv(Filelocation1,names=["Age","Year","axilnodes","survival"])
Hdata[:9]
```

Out[5]:

	Age	Year	axilnodes	survival
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1
5	33	58	10	1
6	33	60	0	1
7	34	59	0	2
8	34	66	9	2

In [40]:
```python
print(Hdata.columns)
print(Hdata.shape)
print(Hdata['survival'].value_counts())
```

```
Index(['Age', 'Year', 'axilnodes', 'survival'], dtype='object')
(306, 4)
1    225
2     81
Name: survival, dtype: int64
```

HIGH LEVEL STATISTICS

1.The Data consits of 305 rows or datapoints and 4 columns or features

2.There are 4 features/variables out of which 3 are independent variables and one named survival which is the last column is dependent variable or class label

3.The columns being Age i.e Age of operation, Year -Year of operation ,axilnodes- no of +ve auxillary nodes and survival - sttus of how long the patient survived

4.The data points of survival does not seem to be balanced

OBJECTIVE

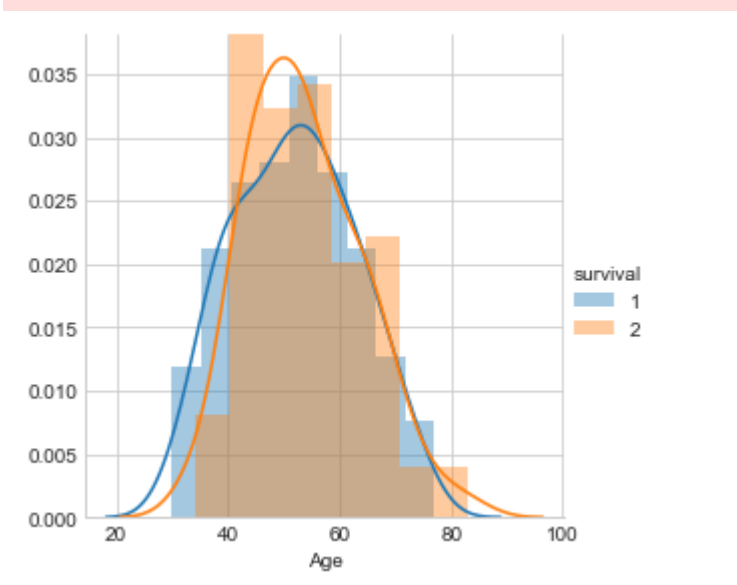
Classifing wether wether the patient survives 5 years or more based on age,year,auxillary nodes

Univariate Analysis:-

Histogram of the independent features :-

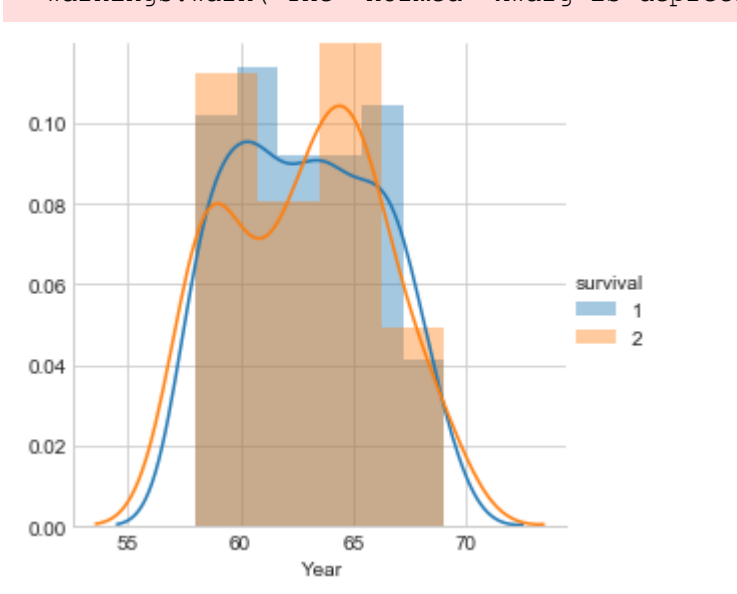
In [6]:
```python
Hdata_1=Hdata[Hdata["survival"]==1]
Hdata_2=Hdata[Hdata["survival"]==2]
plt.close()
sns.set_style("whitegrid");
sns.FacetGrid(Hdata,hue="survival",size=4).map(sns.distplot,'Age').add_legend();
plt.show()
```

```
C:\Users\akars_wj3kiua\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: T
he 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\akars_wj3kiua\Anaconda3\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: T
he 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



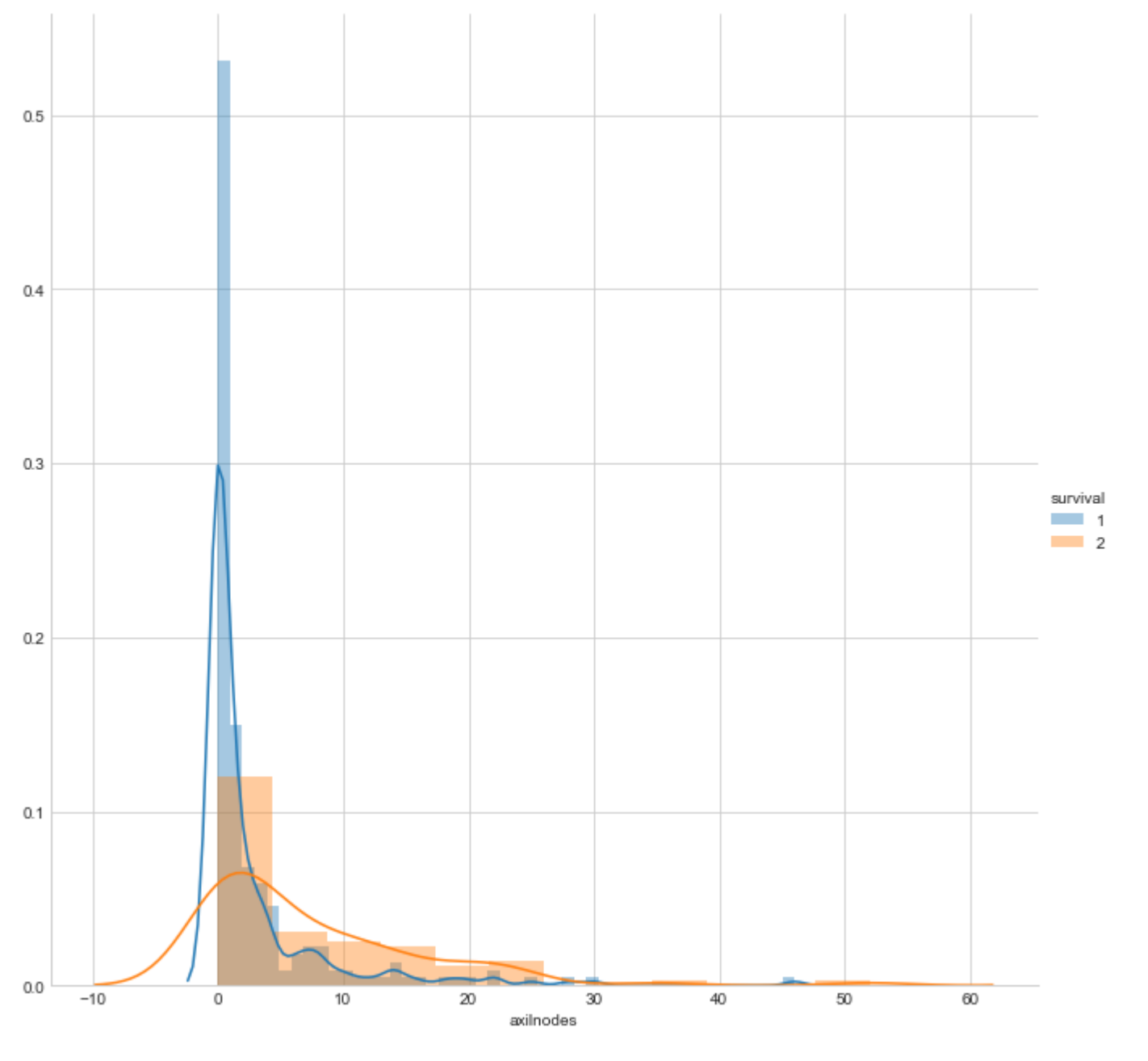
In [50]:
```python
plt.close()
sns.set_style("whitegrid");
sns.FacetGrid(Hdata,hue="survival",size=4).map(sns.distplot,"Year").add_legend()
plt.show()
```

```
C:\Users\A674682\AppData\Local\Continuum\anaconda3\lib\site-packages\matplotlib\axes\_axes.py:64
62: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\A674682\AppData\Local\Continuum\anaconda3\lib\site-packages\matplotlib\axes\_axes.py:64
62: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



In [56]:
```python
plt.close()
sns.set_style("whitegrid");
sns.FacetGrid(Hdata,hue="survival",size=9).map(sns.distplot,"axilnodes").add_legend()
plt.show()
```

```
C:\Users\A674682\AppData\Local\Continuum\anaconda3\lib\site-packages\matplotlib\axes\_axes.py:64
62: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\A674682\AppData\Local\Continuum\anaconda3\lib\site-packages\matplotlib\axes\_axes.py:64
62: UserWarning: The 'normed' kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



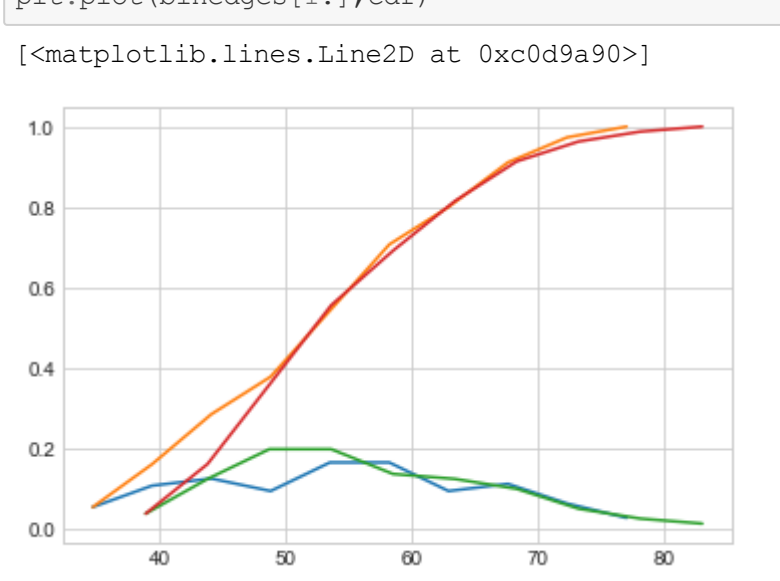
Observation :-

1) We can see that it is not possible to clasify wether the patient survives more than 5 years or less than five yers by looking into each independent variables as Histogram of each feature there is an overlap of the survival status 1 & 2 and hence Single feature alone does not seem useful for clasification

PDF and CDF

In [65]:
```python
count,binedges=np.histogram(Hdata_1["Age"],bins=10,density=True)
pdf=count/(sum(count))
cdf=np.cumsum(pdf)
plt.plot(binedges[1:],pdf)
plt.plot(binedges[1:],cdf)

count,binedges=np.histogram(Hdata_2["Age"],bins=10,density=True)
pdf=count/(sum(count))
cdf=np.cumsum(pdf)
plt.plot(binedges[1:],pdf)
plt.plot(binedges[1:],cdf)
```

Out[65]: `[<matplotlib.lines.Line2D at 0xc0d9a90>]`

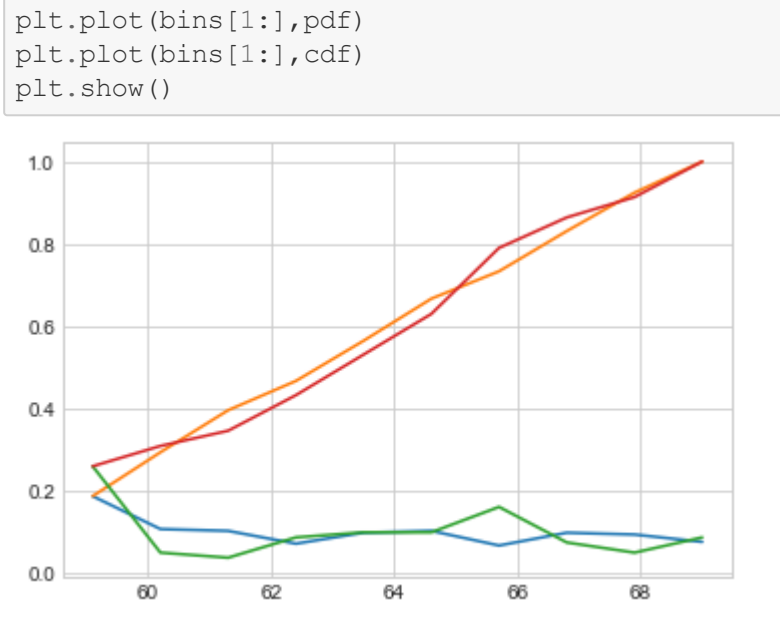


Observation:-

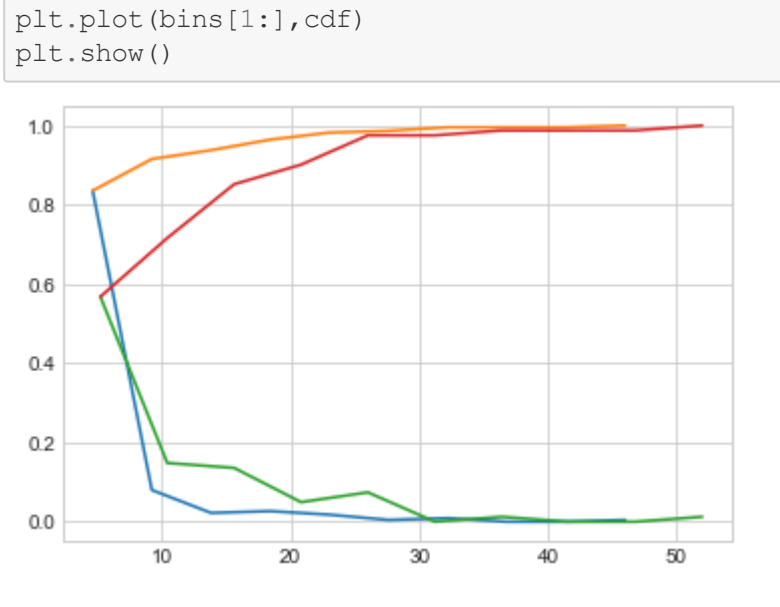
1.CDF of both survival>=5 (i.e survial=1) and survival<5 i.e (survival =2) almost ovrerlap

2.20% of patients who survived 5 years or more were operated at or less than 41 years of age 20% of the patients who didnt survive for 5 years got operated at or less than 45 years of age

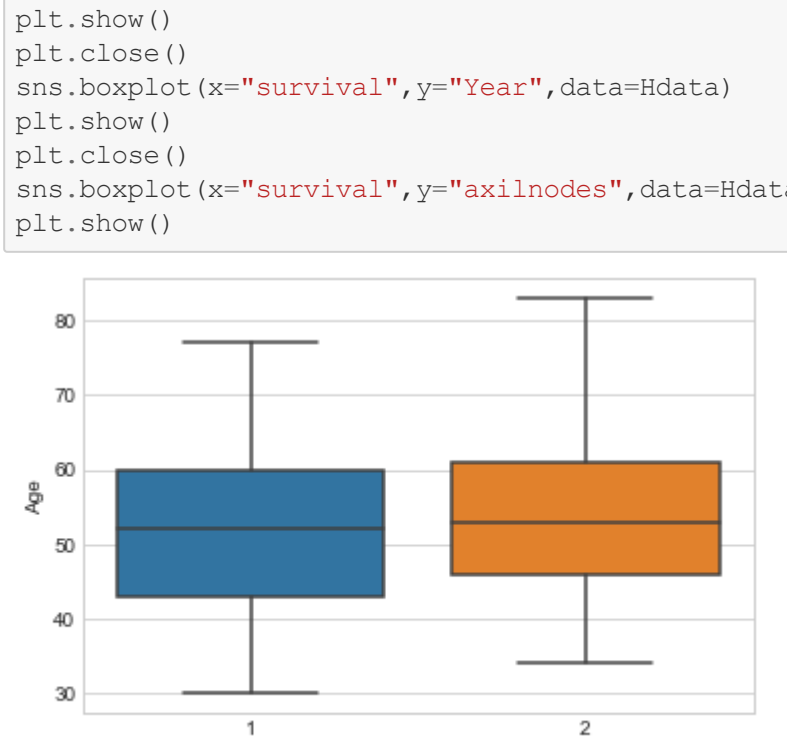
In [70]:
```python
plt.close()
cnt,bins=np.histogram(Hdata_1["Year"],density=True)
pdf=cnt/(sum(cnt))
cdf=np.cumsum(pdf)
plt.plot(bins[1:],pdf)
plt.plot(bins[1:],cdf)


cnt,bins=np.histogram(Hdata_2["Year"],density=True)
pdf=cnt/(sum(cnt))
cdf=np.cumsum(pdf)
plt.plot(bins[1:],pdf)
plt.plot(bins[1:],cdf)
plt.show()
```

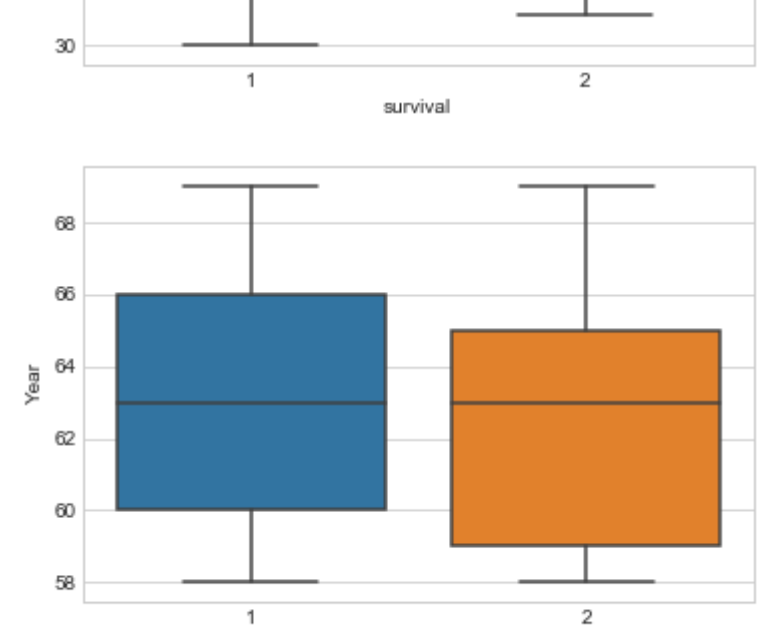


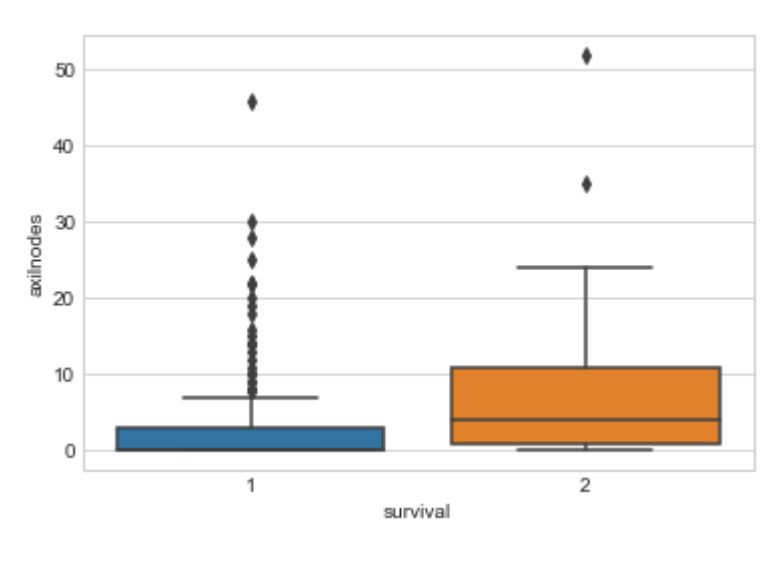
In [71]:
```python
plt.close()
cnt,bins=np.histogram(Hdata_1["axilnodes"],density=True)
pdf=cnt/(sum(cnt))
cdf=np.cumsum(pdf)
plt.plot(bins[1:],pdf)
plt.plot(bins[1:],cdf)


cnt,bins=np.histogram(Hdata_2["axilnodes"],density=True)
pdf=cnt/(sum(cnt))
cdf=np.cumsum(pdf)
plt.plot(bins[1:],pdf)
plt.plot(bins[1:],cdf)
plt.show()
```



In [77]:
```python
plt.close()
sns.boxplot(x="survival",y="Age",data=Hdata)
plt.show()
plt.close()
sns.boxplot(x="survival",y="Year",data=Hdata)
plt.show()
plt.close()
sns.boxplot(x="survival",y="axilnodes",data=Hdata)
plt.show()
```







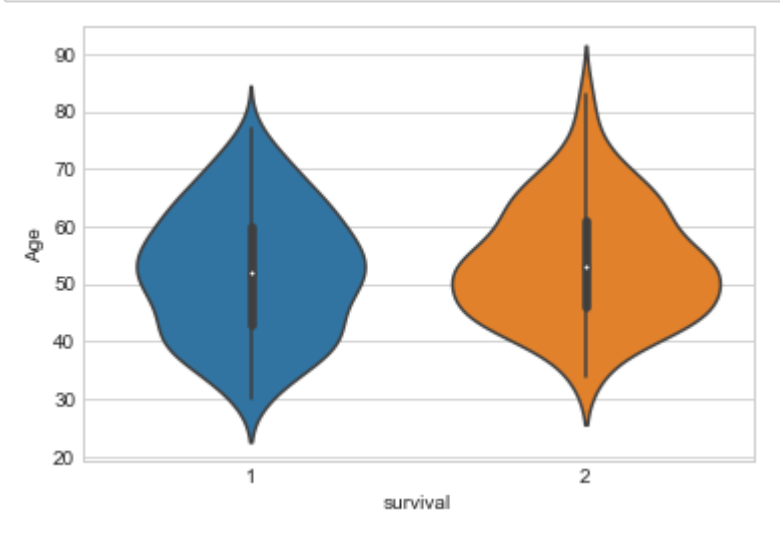
Observations :-

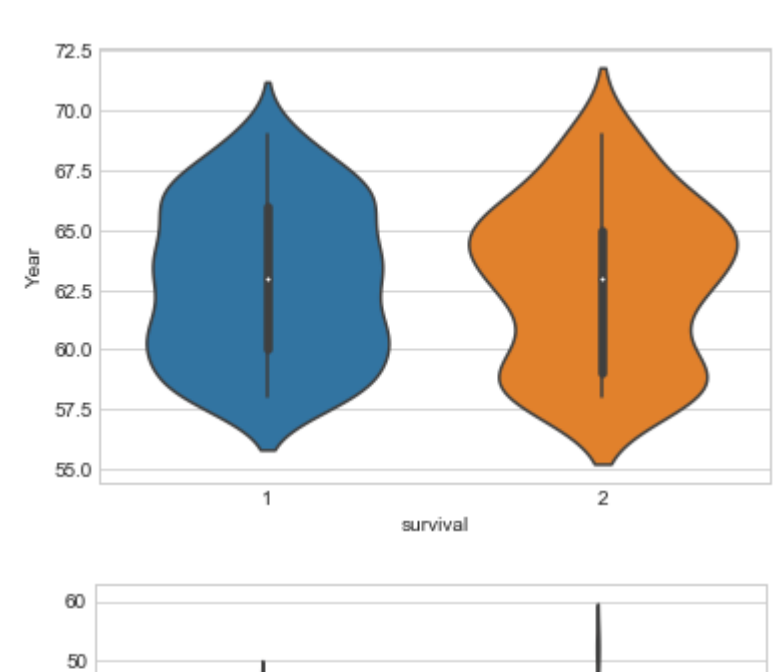
1.Box plots for Age fig(1) are almost similar in terms of ages of 25,50,75 percentile

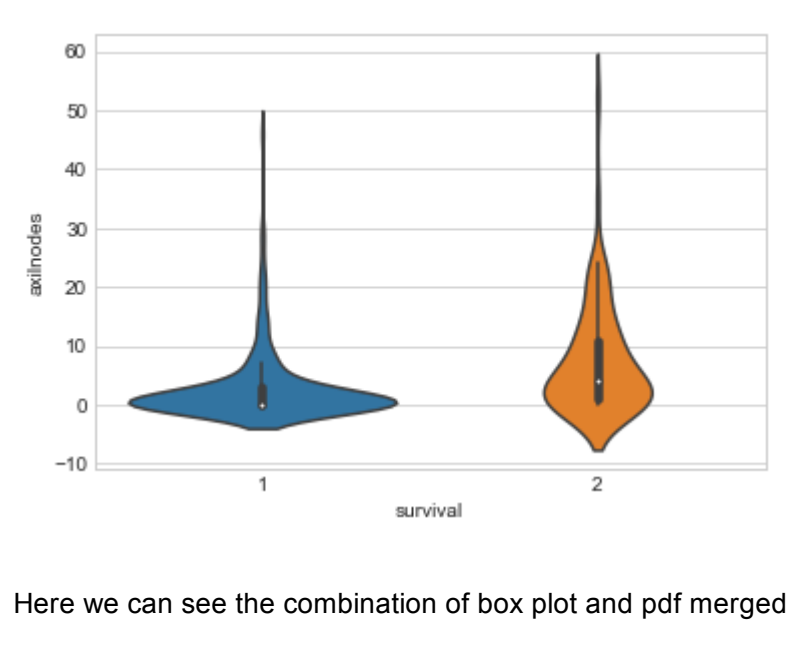
2.Box plots for Year fig(2) are almost similar except 25 % of operations of patients who survived more than 5 years happened in 1960 or later and 75 % occoured before 1966 wheras 25 % of patients who did not survive 5 years were operated before 1959 and 75% before 1965

3.Box plots for auxillary nodes fig(3) 75% of patients who survived for greater than or equal to 5 years had less than 2 positive auxillary nodes wheras 75% of patients who did not survive for 5 years had less than 11 positive auxillary nodes

In [78]:
```python
plt.close()
sns.violinplot(x="survival",y="Age",data=Hdata)
plt.show()
plt.close()
sns.violinplot(x="survival",y="Year",data=Hdata)
plt.show()
plt.close()
sns.violinplot(x="survival",y="axilnodes",data=Hdata)
plt.show()
```







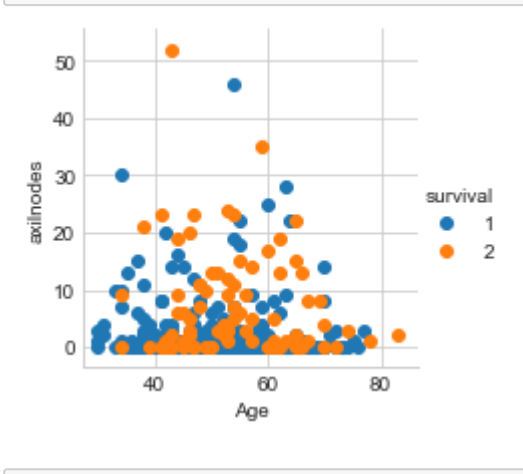
Here we can see the combination of box plot and pdf merged together to form violin plots

Bi-variate analysis

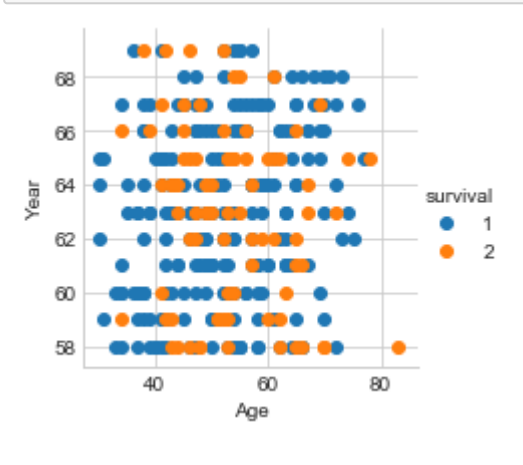
Scatter Plots :-

In [1]:
```python
#Hdata.plot(kind="scatter",x="Age",y="survival")
#plt.show()
```

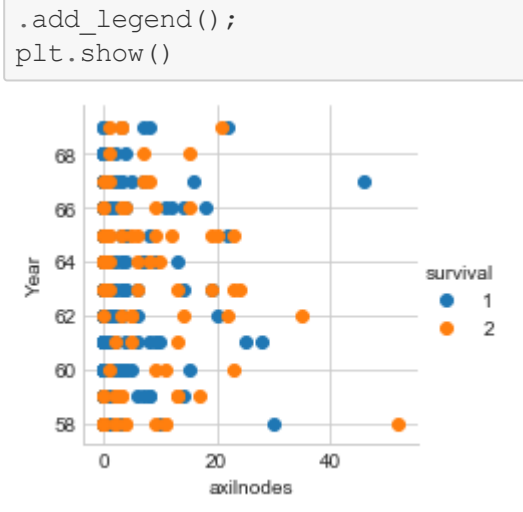
In [22]:
```python
sns.set_style('whitegrid');
sns.FacetGrid(Hdata,hue="survival")\
.map(plt.scatter,"Age","axilnodes")\
.add_legend();
plt.show()
```



In [6]:
```python
sns.set_style('whitegrid');
sns.FacetGrid(Hdata,hue="survival")\
.map(plt.scatter,"Age","Year")\
.add_legend();
plt.show()
```

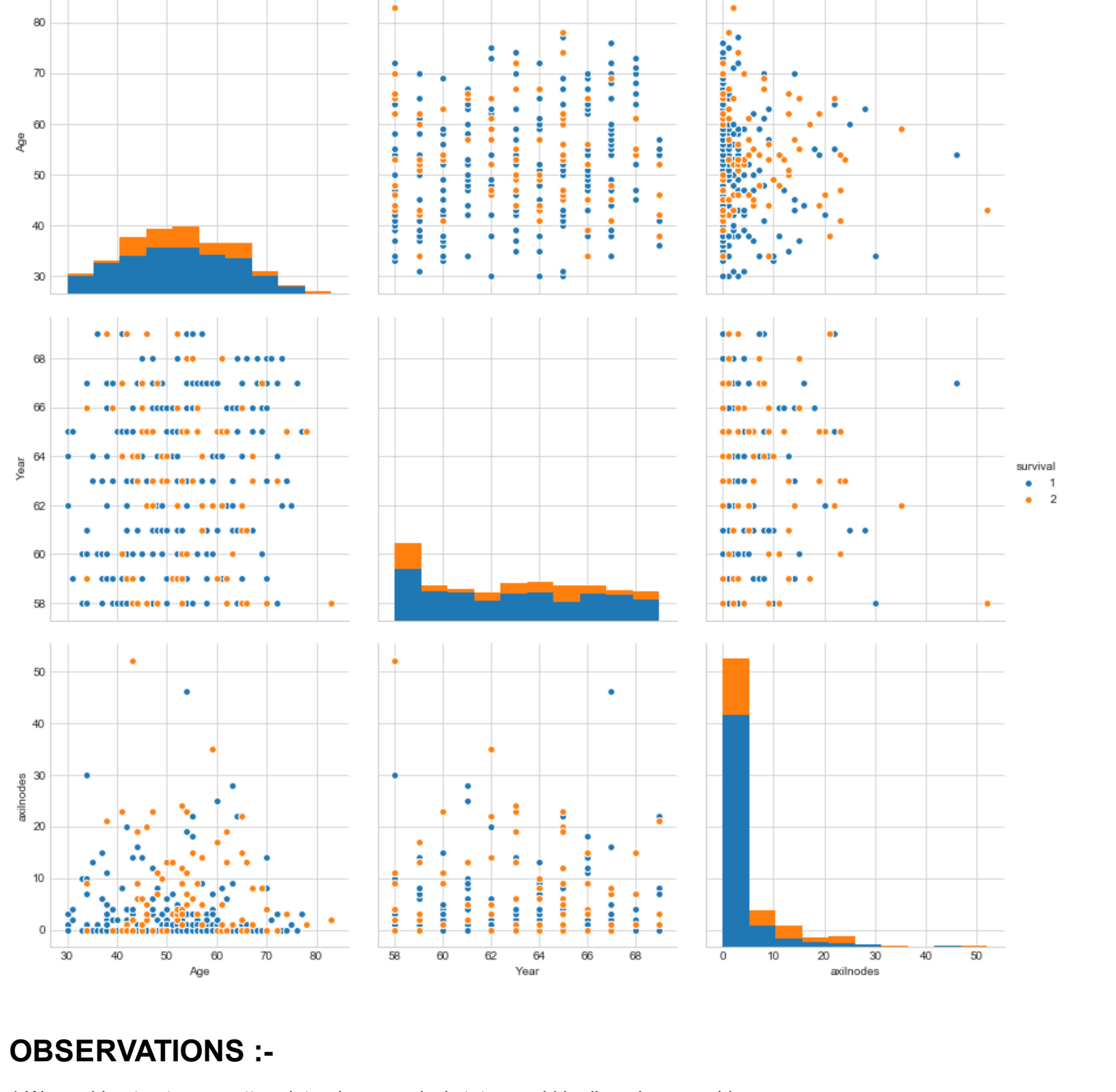


In [8]:
```python
sns.set_style('whitegrid');
sns.FacetGrid(Hdata,hue="survival")\
.map(plt.scatter,"axilnodes","Year")\
.add_legend();
plt.show()
```



PAIR PLOTS:-

In [33]:
```python
plt.close()
sns.pairplot(Hdata,hue="survival",vars=['Age', 'Year', 'axilnodes'],size=4)
plt.show()
```



OBSERVATIONS :-

1.We could not get any scatter plots where survival status could be linearly seperable

2.In case of pair plots the plots of all the features survival staus were higly overlaped and could not be segregated

3.We can conclude that even combination of features was not helpful in classification of patients based on survival

In [27]:
```python
MnAgg5=np.mean(Hdata_1["Age"])
MnAgl5=np.mean(Hdata_2["Age"])
Mnyrg5=np.mean(Hdata_1["Year"])
Mnyrl5=np.mean(Hdata_2["Year"])
Mnaxg5=np.mean(Hdata_1["axilnodes"])
Mnaxl5=np.mean(Hdata_2["axilnodes"])
MdAgg5=np.median(Hdata_1["Age"])
MdAgl5=np.median(Hdata_2["Age"])
Mdyrg5=np.median(Hdata_1["Year"])
Mdyrl5=np.median(Hdata_2["Year"])
Mdaxg5=np.median(Hdata_1["axilnodes"])
Mdaxl5=np.median(Hdata_2["axilnodes"])
```

In [38]:
```python
StdAgg5=np.std(Hdata_1["Age"])
StdAgl5=np.std(Hdata_2["Age"])
Stdyrg5=np.std(Hdata_1["Year"])
Stdyrl5=np.std(Hdata_2["Year"])
Stdaxg5=np.std(Hdata_1["axilnodes"])
Stdaxl5=np.std(Hdata_2["axilnodes"])
```

In [39]:
```python
print("Mean Table :-")
print("------------")
from beautifultable import BeautifulTable
table = BeautifulTable()
table.column_headers = ['Survival Status','Age','Year','No of Auxillary nodes']
table.append_row(['surviving >= 5 years',MnAgg5,Mnyrg5,Mnaxg5])
table.append_row(['surviving <5 years',MnAgl5,Mnyrl5,Mnaxl5])
print(table)
print("")
print("")
print("Median Table :-")
print("------------")
from beautifultable import BeautifulTable
table = BeautifulTable()
table.column_headers = ['Survival Status','Age','Year','No of Auxillary nodes']
table.append_row(['surviving >= 5 years',MdAgg5,Mdyrg5,Mdaxg5])
table.append_row(['surviving <5 years',MdAgl5,Mdyrl5,Mdaxl5])
print(table)
print("Spread Table :-")
print("------------")
from beautifultable import BeautifulTable
table = BeautifulTable()
table.column_headers = ['Survival Status','Age','Year','No of Auxillary nodes']
table.append_row(['surviving >= 5 years',StdAgg5,Stdyrg5,Stdaxg5])
table.append_row(['surviving <5 years',StdAgl5,Stdyrl5,Stdaxl5])
print(table)
```

```
Mean Table :-
------------
+----------------------------+--------+--------+-----------------------------+
|      Survival Status       |  Age   |  Year  |    No of Auxillary nodes    |
+----------------------------+--------+--------+-----------------------------+
|    surviving >= 5 years    | 52.018 | 62.862 |            2.791            |
+----------------------------+--------+--------+-----------------------------+
|     surviving <5 years     | 53.679 | 62.827 |            7.457            |
+----------------------------+--------+--------+-----------------------------+


Median Table :-
------------
+------------------------------+------+------+-----------------------------+
|        Survival Status       | Age  | Year |    No of Auxillary nodes    |
+------------------------------+------+------+-----------------------------+
|     surviving >= 5 years     | 52.0 | 63.0 |             0.0             |
+------------------------------+------+------+-----------------------------+
|      surviving <5 years      | 53.0 | 63.0 |             4.0             |
+------------------------------+------+------+-----------------------------+
Spread Table :-
------------
+----------------------------+--------+-------+-----------------------------+
|      Survival Status       |  Age   | Year  |    No of Auxillary nodes    |
+----------------------------+--------+-------+-----------------------------+
|    surviving >= 5 years    | 10.988 | 3.216 |            5.857            |
+----------------------------+--------+-------+-----------------------------+
|     surviving <5 years     | 10.104 | 3.321 |            9.129            |
+----------------------------+--------+-------+-----------------------------+
```

Observations:-

1. Central tendancy of Age and Year are quite similar for both the survival categories
2. There is a shift in central tendacies of No of positive aux nodes of both survival categories
3. Although we can see shift in central tendancies for No of Auxillary nodes but this wont be suitable for classification of survival status as seen from the spread table above <5 years has a larger spread and hence as also seen by box plots more than 50 % of the data overlap and there is no linear biforcation visible