Attribute info

1.age: age of patient

2.op-year:year of the operation

3.axil-nodes: no of axilary nodes detected

4.survived: survival status

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
data=pd.read_csv("haberman.csv")   #read csvv file into data
```

```
#data points and feaures
print(data.shape)
```

```
(305, 4)
```

conclusion

1. There are 4 features with 305 data-points/observation for each feature.

```
#the data set columns are named
data.columns=['age','op-year','axil-nodes','survived'] # column named
#data[data==2]=0 # 2 to 0
data.tail()
data.columns
```

```
Index(['age', 'op-year', 'axil-nodes', 'survived'], dtype='object')
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 305 entries, 0 to 304
Data columns (total 4 columns):
age           305 non-null int64
op-year       305 non-null int64
axil-nodes    305 non-null int64
survived      305 non-null int64
```

conclusion

the values are not null.ie data set is already cleaned, no missing values

```
#patiens survived and not
print(data['survived'].value_counts())
# print(data['op-year'].value_counts())
```

```
1    224
2     81
Name: survived, dtype: int64
```

conclusions

1 for survived and 2 for not survived

224 patients have survival status 1, who lived for more than 5 year

81 patients have survival status 2, who lived for less than 5 year

```
#percentage of people survived
x=str(225/304)
y=str(81/304)
print("people survived is in  % is "+x)
print("people not survived is in  % is "+y)
```

```
people survived is in  % is 0.7401315789473685
people not survived is in  % is 0.26644736842105265
```

```
data.describe()
```

	age	op-year	axil-nodes	survived
count	305.000000	305.000000	305.000000	305.000000
mean	52.531148	62.849180	4.036066	1.265574
std	10.744024	3.254078	7.199370	0.442364
min	30.000000	58.000000	0.000000	1.000000
25%	44.000000	60.000000	0.000000	1.000000
50%	52.000000	63.000000	1.000000	1.000000
75%	61.000000	66.000000	4.000000	2.000000

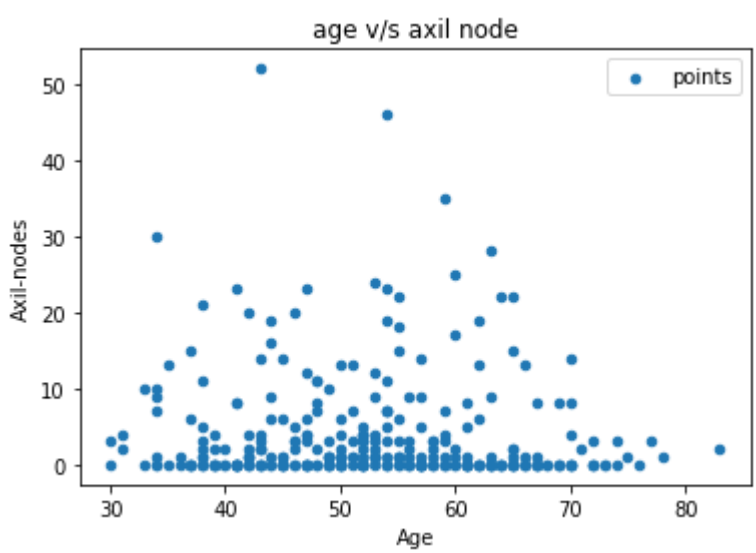
conclusions

74% chances to survive after operation and 26% to not survive

25 % of women had 0 axilary nodes

75% of people had 4 axilary nodes

```
data.plot(kind='scatter',x='age',y='axil-nodes',label="points")
plt.xlabel("Age")
plt.ylabel("Axil-nodes")
plt.title("age v/s axil node")
plt.legend()
plt.show()
```

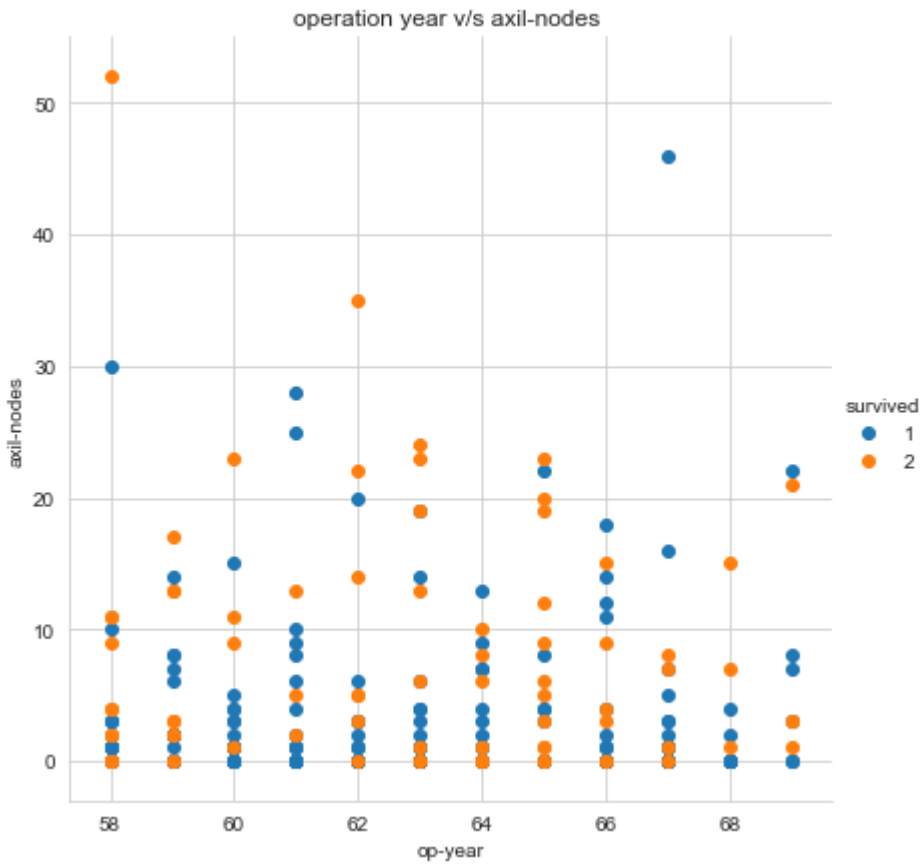



```
sns.set_style("whitegrid")
sns.pairplot(data,hue='survived',height=3,x_vars={'age','op-year','axil-nodes'},y_vars={'age'
plt.show()
```

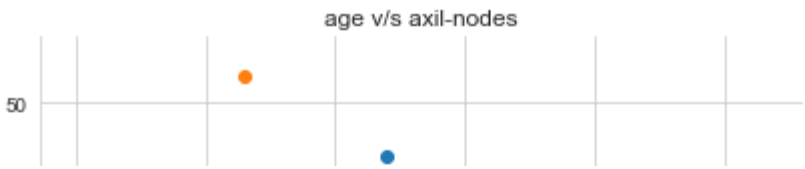
conclusion

this is an unbalance dataset

```
sns.set_style("whitegrid")
sns.FacetGrid(data,hue="survived",height=6).map(plt.scatter,"op-year","axil-nodes").add_legen
plt.title("operation year v/s axil-nodes ")
plt.xlabel("op-year")
plt.ylabel("axil-nodes")
plt.show()
```

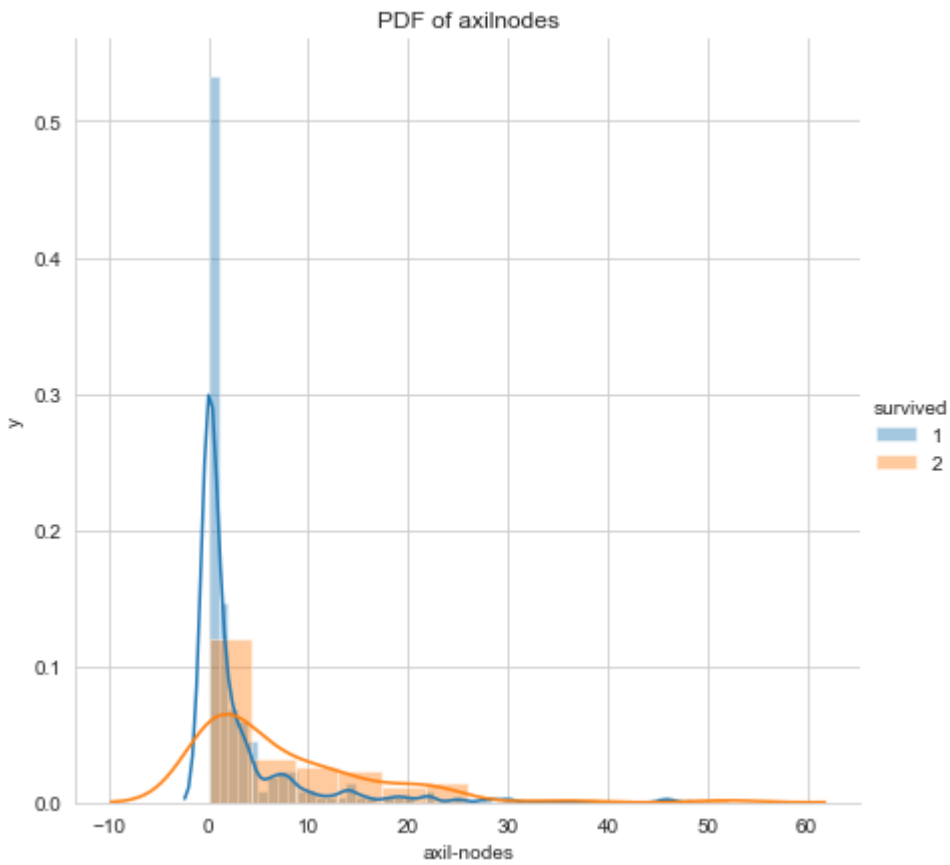



```
sns.set_style("whitegrid")
sns.FacetGrid(data,hue="survived",height=6).map(plt.scatter,"age","axil-nodes").add_legend();
plt.title("age v/s axil-nodes")
plt.xlabel("age")
plt.ylabel("axil-node")
plt.show()
```



PDF for columns

```
sns.FacetGrid(data,hue='survived',height=6)\
    .map(sns.distplot,'axil-nodes')\
    .add_legend()
plt.xlabel("axil-nodes")
plt.ylabel("y")
plt.title("PDF of axilnodes")
plt.show()
```

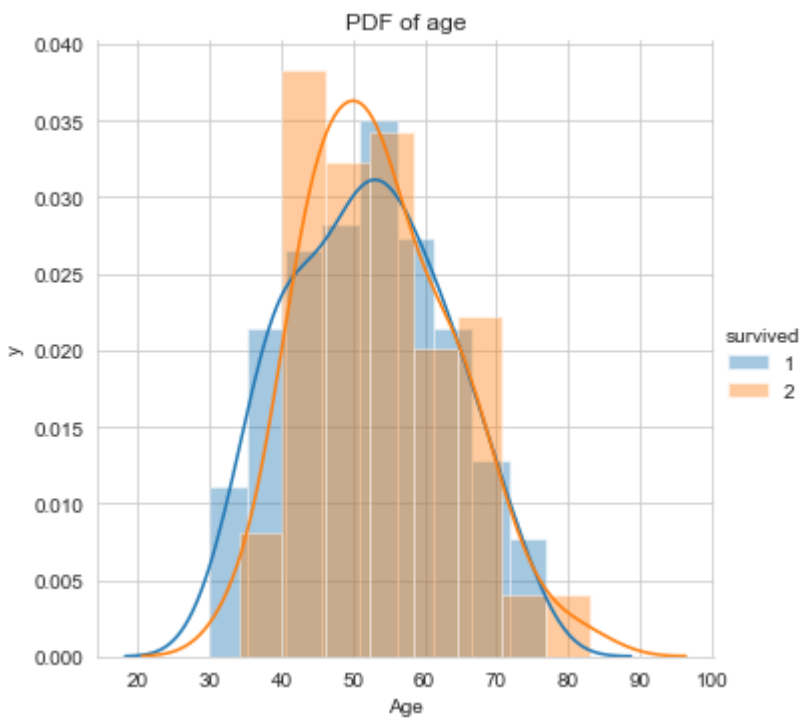


conclusion

"axilary nodes " is the only feature which shows some useful insights as there is a differnce between the distribution of both

classes/labels. More patients survied who have zero number of axilary nodes

```
plt.close();
sns.FacetGrid(data,hue='survived',height=5)\
    .map(sns.distplot,"age")\
    .add_legend()
plt.xlabel("Age")
plt.ylabel("y")
plt.title("PDF of age")
plt.show()
```



conclusions

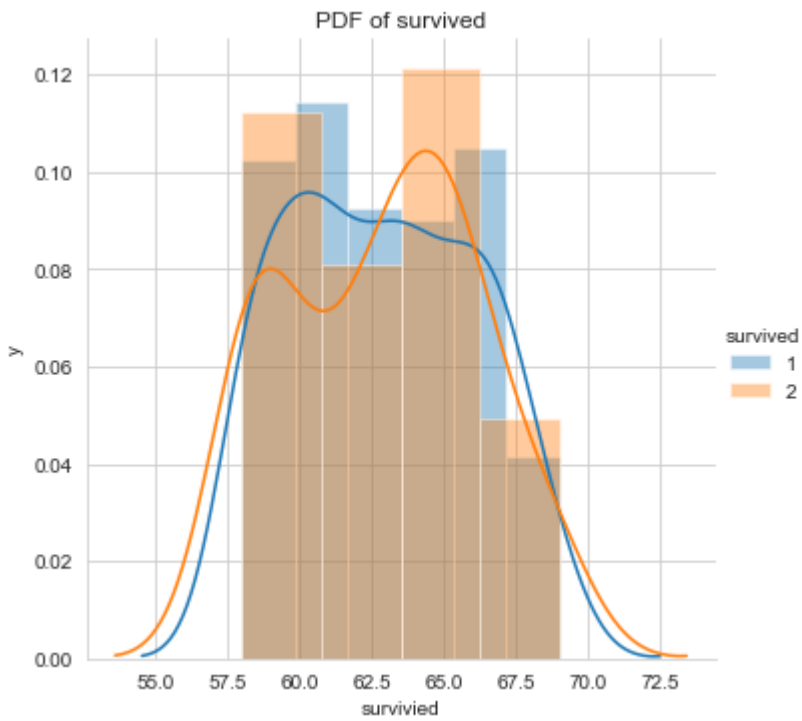
the graphs are overlaping with eachother

the people in between age 40-60 are more likely to die

the people less than age 40 more likely to survive

```
plt.close();
sns.FacetGrid(data,hue="survived",height=5).map(sns.distplot,"op-year").add_legend()
plt.xlabel("survivied")
```

```
plt.ylabel("y")
plt.title("PDF of survived")
plt.show()
```



conclusions

operation year in between 1957 and 1967 many people not survived

the age and operation year is not useful features to find insight , because they are not well separated ,lived and died people insights are overlapped

the graph is not well separated , so the univarient analysis is not found to be useful

```
lived=data.loc[data['survived']==1]
lived.describe()
```

	age	op-year	axil-nodes	survived
count	224.000000	224.000000	224.000000	224.0
mean	52.116071	62.857143	2.799107	1.0
std	10.937446	3.229231	5.882237	0.0
min	30.000000	58.000000	0.000000	1.0
25%	43.000000	60.000000	0.000000	1.0
50%	[illegible]	[illegible]	[illegible]	[illegible]

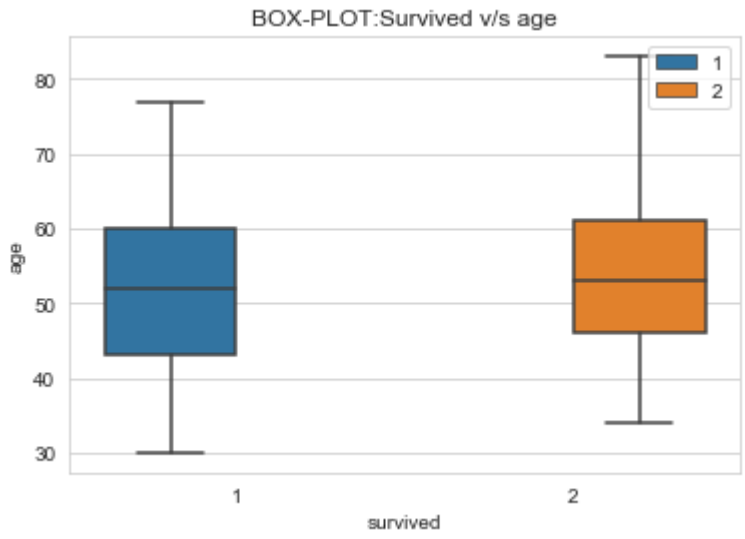
```
died=data.loc[data['survived']==2]
died.describe()
```

	age	op-year	axil-nodes	survived
count	81.000000	81.000000	81.000000	81.0
mean	53.679012	62.827160	7.456790	2.0
std	10.167137	3.342118	9.185654	0.0
min	34.000000	58.000000	0.000000	2.0
25%	46.000000	59.000000	1.000000	2.0
50%	53.000000	63.000000	4.000000	2.0
75%	61.000000	65.000000	11.000000	2.0
max	83.000000	69.000000	52.000000	2.0

conclusion

Based on above two outputs we can get an trend that 75 % women that lived had less than 3 positive axillary nodes detected . While women that died , 50% Women had more than 4 positive axillary nodes detected .

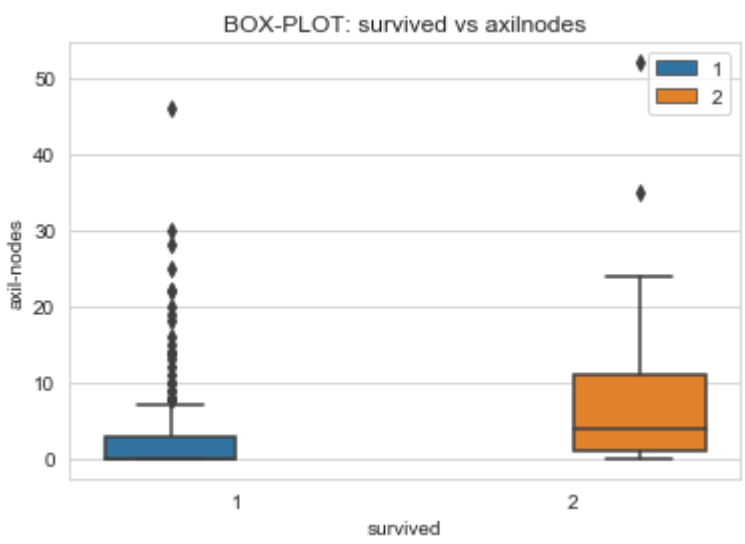
```
plt.close();
sns.boxplot(x='survived',y='age',hue="survived",data=data)
plt.xlabel("survived")
plt.ylabel("age")
plt.title("BOX-PLOT:Survived v/s age")
plt.legend(loc=1)
plt.show()
```



conclusion

almost 95% of age overlaps so age is not sufficient for the analysis

```
plt.close();
sns.boxplot('survived','axil-nodes',hue="survived",data=data)
plt.xlabel("survived")
plt.ylabel("axil-nodes")
plt.title("BOX-PLOT: survived vs axilnodes")
plt.legend(loc=0)
plt.show()
```

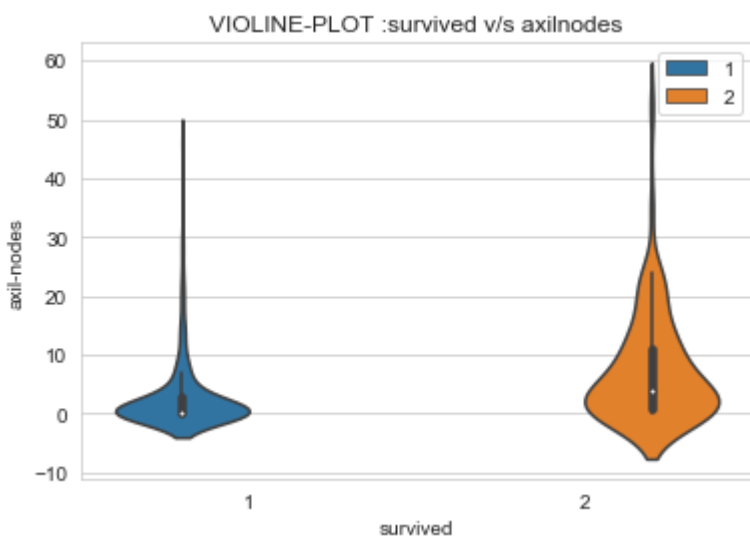


conclusions

the first boxplot have outliers

the box plot also shows that there is a 60-65% of misclassification

```
plt.close();
sns.violinplot(x='survived',y='axil-nodes',data=data,hue='survived',height=4)
plt.xlabel("survived")
plt.ylabel("axil-nodes")
plt.title("VIOLINE-PLOT :survived v/s axilnodes")
plt.legend(loc=0)
plt.show()
```



conclusion

survival person have axilnode =0

if the axilnode is increses there is less chance to survive

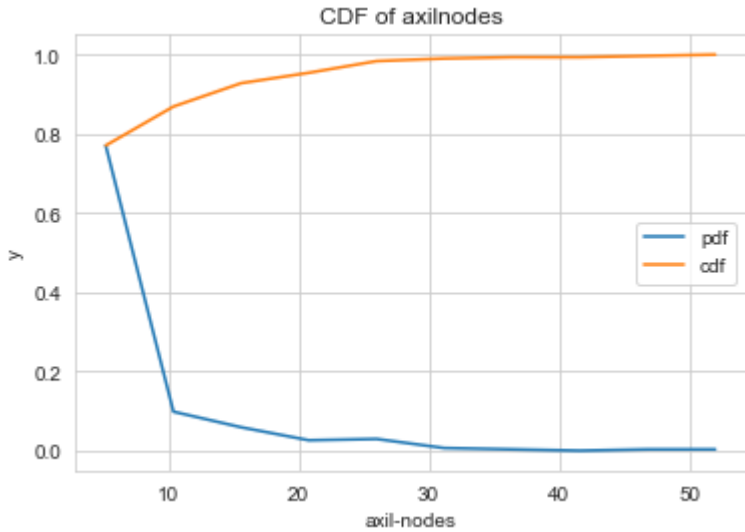
```
counts, bin_edges = np.histogram(data['axil-nodes'], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges)

#compute CDF
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label="pdf")
plt.plot(bin_edges[1:], cdf,label="cdf")
plt.xlabel("axil-nodes")

plt.ylabel("y")
```

```
plt.title("CDF of axilnodes")
plt.legend()
plt.show();
```

```
[0.7704918  0.09836066 0.05901639 0.02622951 0.0295082  0.00655738
 0.00327869 0.         0.00327869 0.00327869]
[ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
```



conclusion

if the axilnode is less more chance to survive

if axilnode is more less chance to survive