3. Plotting for Exploratory data analysis (EDA)

(3.12) Exercise:

1. Download Haberman Cancer Survival dataset from Kaggle. You may have to create a Kaggle account to donwload data. (https://www.kaggle.com/gilsousa/habermans-survival-data-set)
2. Perform a similar analysis as above on this dataset with the following sections:

- High level statistics of the dataset: number of points, numer of features, number of classes, data-points per class.
- Explain our objective.
- Perform Univaraite analysis(PDF, CDF, Boxplot, Voilin plots) to understand which features are useful towards classification.
- Perform Bi-variate analysis (scatter plots, pair-plots) to see if combinations of features are useful in classfication.
- Write your observations in english as crisply and unambigously as possible. Always quantify your results.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

#Load Haberman.csv into a pandas dataFrame.

haberman = pd.read_csv("haberman.csv")
haberman["status"] = haberman["status"].replace(1, "the patient
survived 5 years or longer") # replace the status 1/0 to its
respective meaning for more readability
haberman["status"] = haberman["status"].replace(2, "the patient died
within 5 year")
haberman
```

	age	year	nodes	status
0	30	64	1	the patient survived 5 years or longer
1	30	62	3	the patient survived 5 years or longer
2	30	65	0	the patient survived 5 years or longer
3	31	59	2	the patient survived 5 years or longer
4	31	65	4	the patient survived 5 years or longer
5	33	58	10	the patient survived 5 years or longer
6	33	60	0	the patient survived 5 years or longer
7	34	59	0	the patient died within 5 year
8	34	66	9	the patient died within 5 year
9	34	58	30	the patient survived 5 years or longer
10	34	60	1	the patient survived 5 years or longer
11	34	61	10	the patient survived 5 years or longer

12	34	67	7	the patient survived 5 years or longer
13	34	60	0	the patient survived 5 years or longer
14	35	64	13	the patient survived 5 years or longer
15	35	63	0	the patient survived 5 years or longer
16	36	60	1	the patient survived 5 years or longer
17	36	69	0	the patient survived 5 years or longer
18	37	60	0	the patient survived 5 years or longer
19	37	63	0	the patient survived 5 years or longer
20	37	58	0	the patient survived 5 years or longer
21	37	59	6	the patient survived 5 years or longer
22	37	60	15	the patient survived 5 years or longer
23	37	63	0	the patient survived 5 years or longer
24	38	69	21	the patient died within 5 year
25	38	59	2	the patient survived 5 years or longer
26	38	60	0	the patient survived 5 years or longer
27	38	60	0	the patient survived 5 years or longer
28	38	62	3	the patient survived 5 years or longer
29	38	64	1	the patient survived 5 years or longer
..	...	...	...	...
276	67	66	0	the patient survived 5 years or longer
277	67	61	0	the patient survived 5 years or longer
278	67	65	0	the patient survived 5 years or longer
279	68	67	0	the patient survived 5 years or longer
280	68	68	0	the patient survived 5 years or longer
281	69	67	8	the patient died within 5 year
282	69	60	0	the patient survived 5 years or longer
283	69	65	0	the patient survived 5 years or longer
284	69	66	0	the patient survived 5 years or longer
285	70	58	0	the patient died within 5 year
286	70	58	4	the patient died within 5 year
287	70	66	14	the patient survived 5 years or longer
288	70	67	0	the patient survived 5 years or longer
289	70	68	0	the patient survived 5 years or longer
290	70	59	8	the patient survived 5 years or longer
291	70	63	0	the patient survived 5 years or longer
292	71	68	2	the patient survived 5 years or longer
293	72	63	0	the patient died within 5 year
294	72	58	0	the patient survived 5 years or longer
295	72	64	0	the patient survived 5 years or longer
296	72	67	3	the patient survived 5 years or longer
297	73	62	0	the patient survived 5 years or longer
298	73	68	0	the patient survived 5 years or longer
299	74	65	3	the patient died within 5 year
300	74	63	0	the patient survived 5 years or longer
301	75	62	1	the patient survived 5 years or longer
302	76	67	0	the patient survived 5 years or longer
303	77	65	3	the patient survived 5 years or longer
304	78	65	1	the patient died within 5 year
305	83	58	2	the patient died within 5 year

```
[306 rows x 4 columns]
```

```
print(haberman.shape)
```

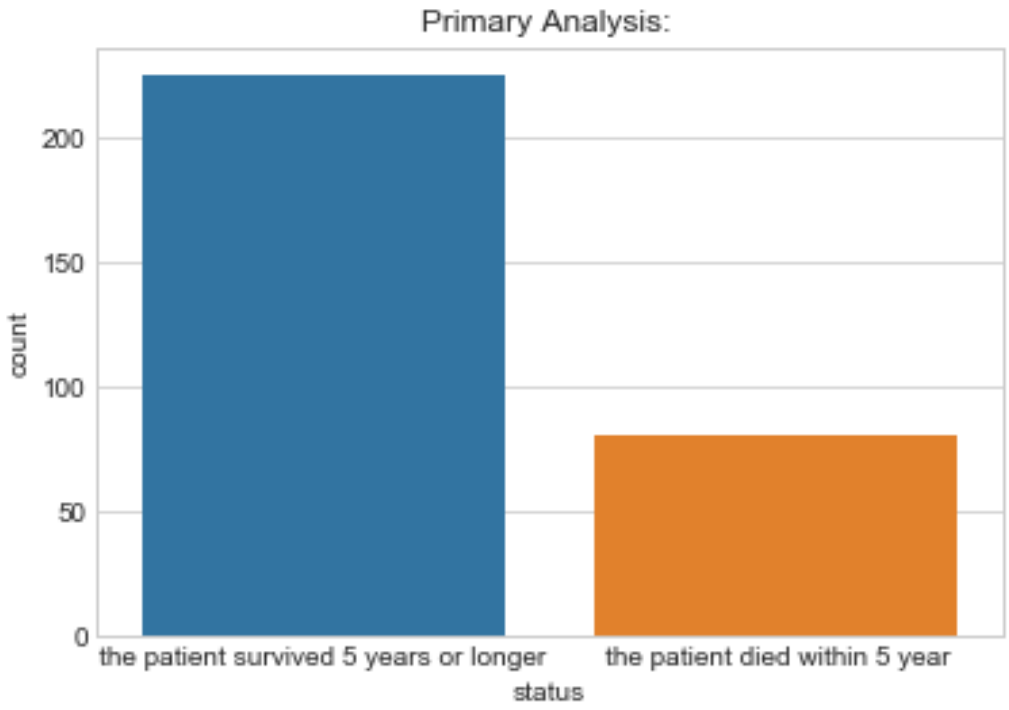
```
(306, 4)
```

```
print(haberman.columns)
```

```
Index(['age', 'year', 'nodes', 'status'], dtype='object')
```

```
#Primary analysis: No of people survived/died after 5 years of
operation

survived = list(haberman[haberman.status == "the patient survived 5
years or longer"] ["status"].value_counts())
print("no of people survived 5 years or longer:",survived)

died =  list(haberman[haberman.status == "the patient died within 5
year"] ["status"].value_counts())
print("no of people died within 5 years:",died)

sns.countplot(x ='status', data = haberman)
fig = plt.gcf() #getcurrentfigure
fig.set_size_inches(6,4)
plt.title('Primary Analysis: ')
plt.show()
```

```
no of people survived 5 years or longer: [225]
no of people died within 5 years: [81]
```



Observation:

```
        225/306 survived 5 years or longer post surgery for breast
cancer.
        81/306 died within 5 years of surgery
```

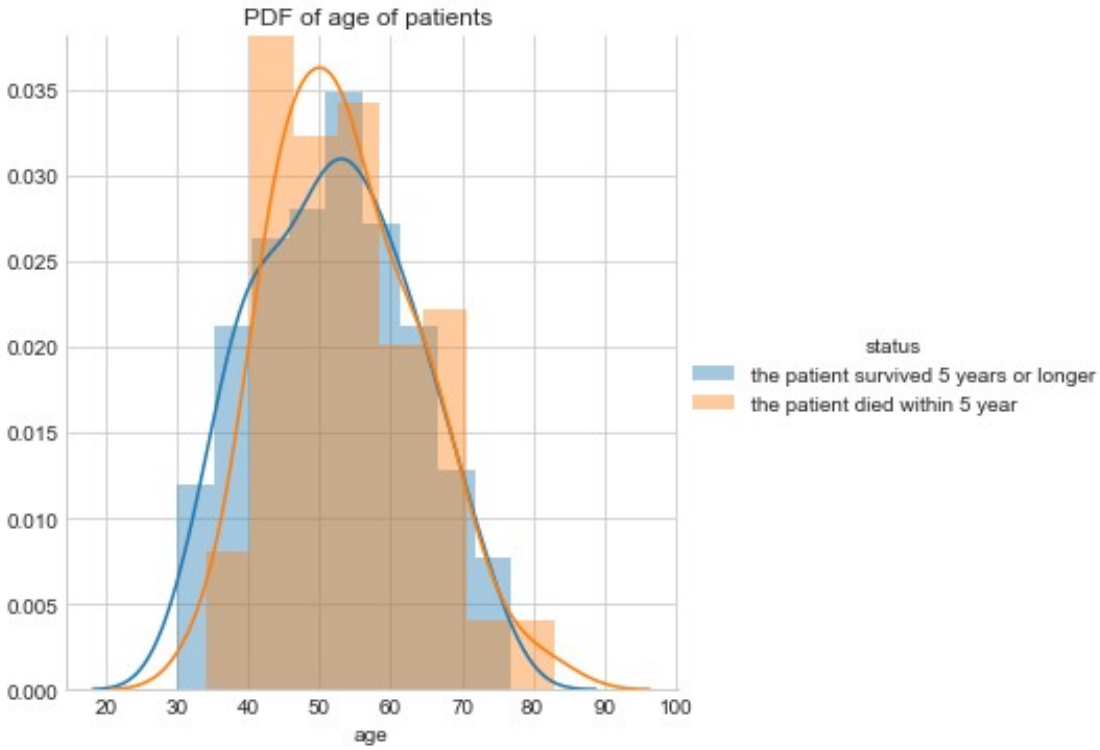
Objective:

To understand the haberman data set and draw insights out of it by performing univariant, bivariant and multivaiant analysis.Also, to determine which factors to be considered primarly before surgery to get better survival rate

```
#Univariant Analysis

#Understanding which age groups survived over the operations from
given dataset

sns.FacetGrid(haberman, hue="status", size=5)\
    .map(sns.distplot, "age")\
    .add_legend()

plt.title('PDF of age of patients ')
plt.show()
```

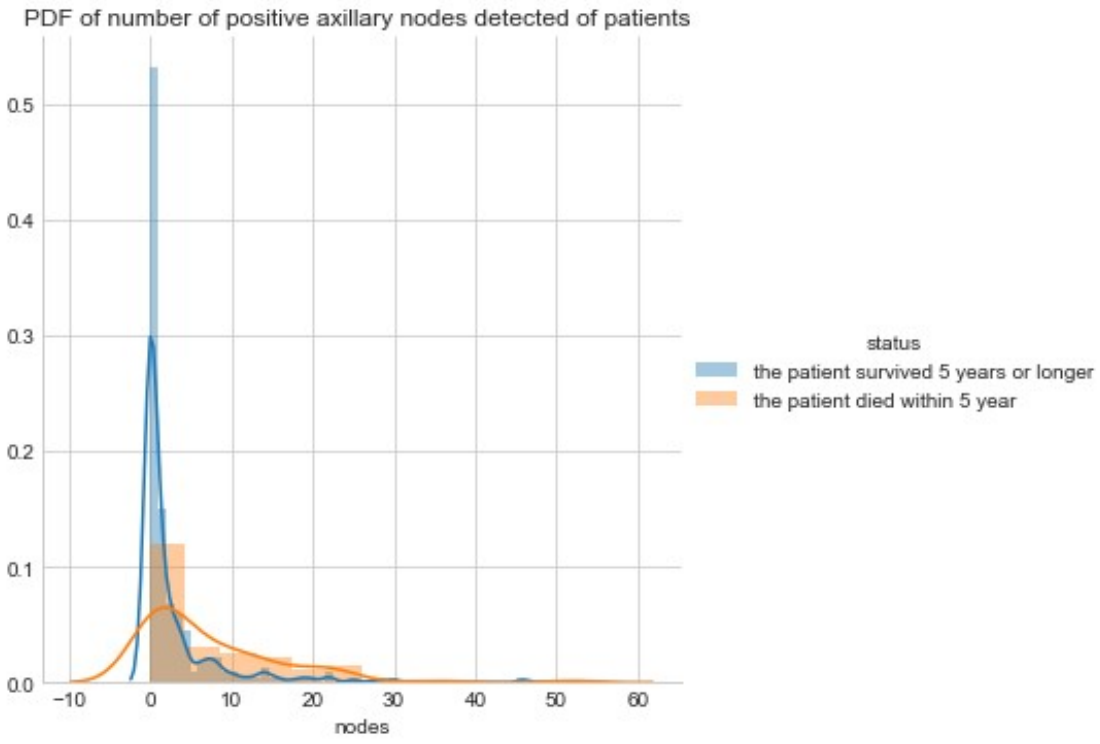
```
C:\Users\akash.ragothu\AppData\Local\Continuum\anaconda3\lib\site-
packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed'
kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

```
C:\Users\akash.ragothu\AppData\Local\Continuum\anaconda3\lib\site-
packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed'
kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```




```
#Understanding number of positive axillary nodes detected vs survived
from given dataset

sns.FacetGrid(haberman, hue="status", size=5)\
    .map(sns.distplot, "nodes")\
    .add_legend()

plt.title('PDF of number of positive axillary nodes detected of
patients ')
plt.show()
```

```
C:\Users\akash.ragothu\AppData\Local\Continuum\anaconda3\lib\site-
packages\matplotlib\axes\_axes.py:6462: 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\akash.ragothu\AppData\Local\Continuum\anaconda3\lib\site-
packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed'
kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



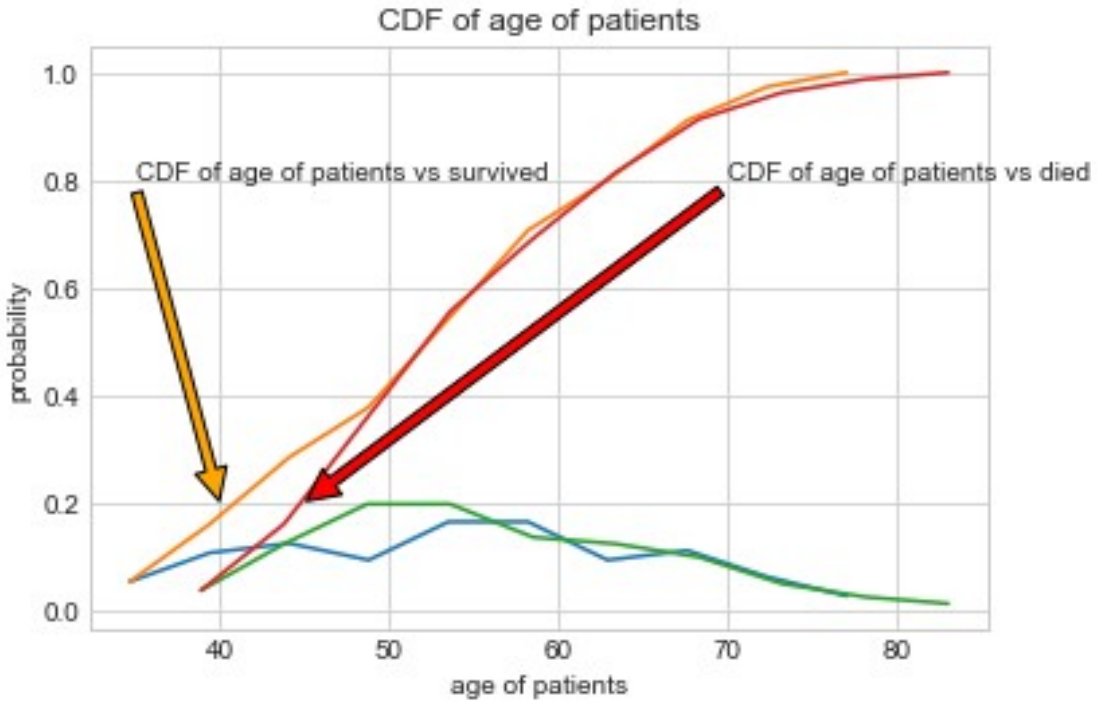
Observation:

PDF of age of patients vs survived is ploted but no much inference made out of it as PDFs are interfering to eachother closely.

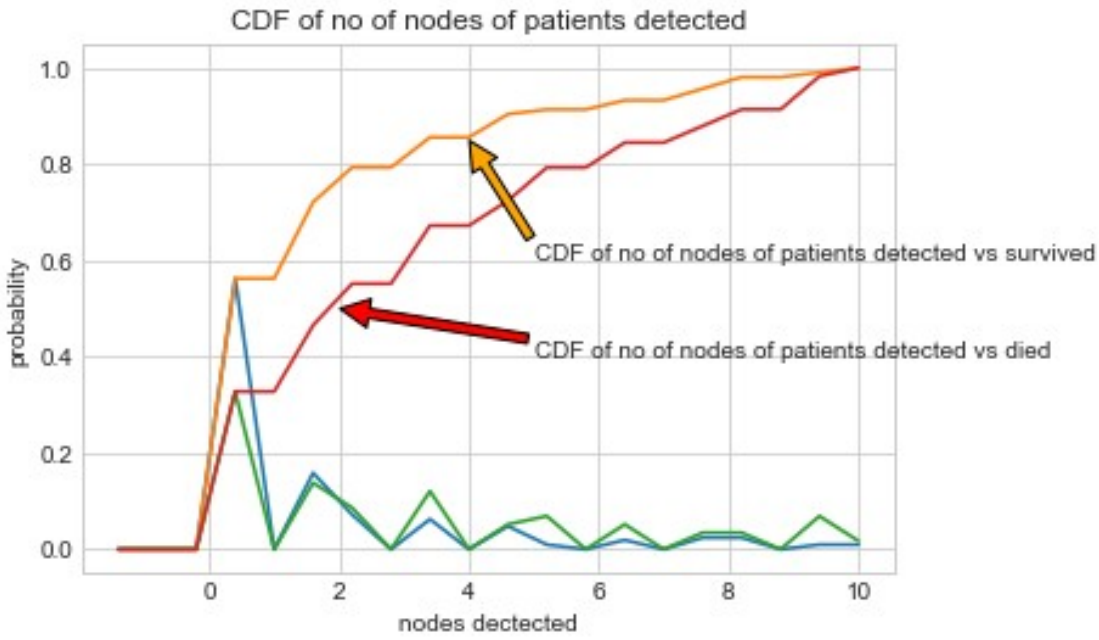
PDF of number of positive axillary nodes detected vs survived is ploted but no much inference made out of it as PDFs are interfering to eachother closely.

```
#source for annotations:
https://jakevdp.github.io/PythonDataScienceHandbook/04.09-text-and-
annotation.html#:~:text=Instead%2C%20I%27d%20suggest%20using%20the
%20plt.annotate%20%28%29%20function.,the%20arrowprops%20dictionary%2C
%20which%20has%20numerous%20options%20available.


survived = haberman[ haberman.status == "the patient survived 5 years
or longer"]
died = haberman[ haberman.status == "the patient died within 5 year"]

# CDF of age of patients vs survived

counts, bin_edges = np.histogram(survived["age"], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)

fig,ax = plt.subplots()
```

```
ax.plot(bin_edges[1:],pdf)
ax.plot(bin_edges[1:], cdf)

# CDF of age of patients vs died

counts, bin_edges = np.histogram(died["age"], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)

ax.plot(bin_edges[1:],pdf)
ax.plot(bin_edges[1:], cdf)

#annotate for better readability

ax.annotate('CDF of age of patients vs survived', xy=(40, 0.2),
xytext=(35, 0.8),
            arrowprops=dict(facecolor='orange', shrink=0.005))
ax.annotate('CDF of age of patients vs died', xy=(45, 0.2),
xytext=(70, 0.8),
            arrowprops=dict(facecolor='red', shrink=0.005))
ax.set_title("CDF of age of patients")
ax.set_xlabel("age of patients")
ax.set_ylabel("probability")

plt.show();
```



```
# CDF of Number of positive axillary nodes detected vs survived

counts, bin_edges = np.histogram(survived["nodes"], bins=20,range = (-
2,10),
                                 density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)

fig,ax = plt.subplots()
ax.plot(bin_edges[1:],pdf)
ax.plot(bin_edges[1:], cdf)

# CDF of Number of positive axillary nodes detected vs died

counts, bin_edges = np.histogram(died["nodes"], bins=20, range = (-
2,10),
                                  density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)

ax.plot(bin_edges[1:],pdf)
ax.plot(bin_edges[1:], cdf)

#annotate for better readability

ax.annotate('CDF of no of nodes of patients detected vs survived',
xy=(4, 0.85), xytext=(5, 0.6),
            arrowprops=dict(facecolor='orange', shrink=0.005))
ax.annotate('CDF of no of nodes of patients detected vs died', xy=(2,
0.5), xytext=(5, 0.4),
            arrowprops=dict(facecolor='red', shrink=0.005))
ax.set_title("CDF of no of nodes of patients detected")
ax.set_xlabel("nodes dectected")
ax.set_ylabel("probability")

plt.show();
```

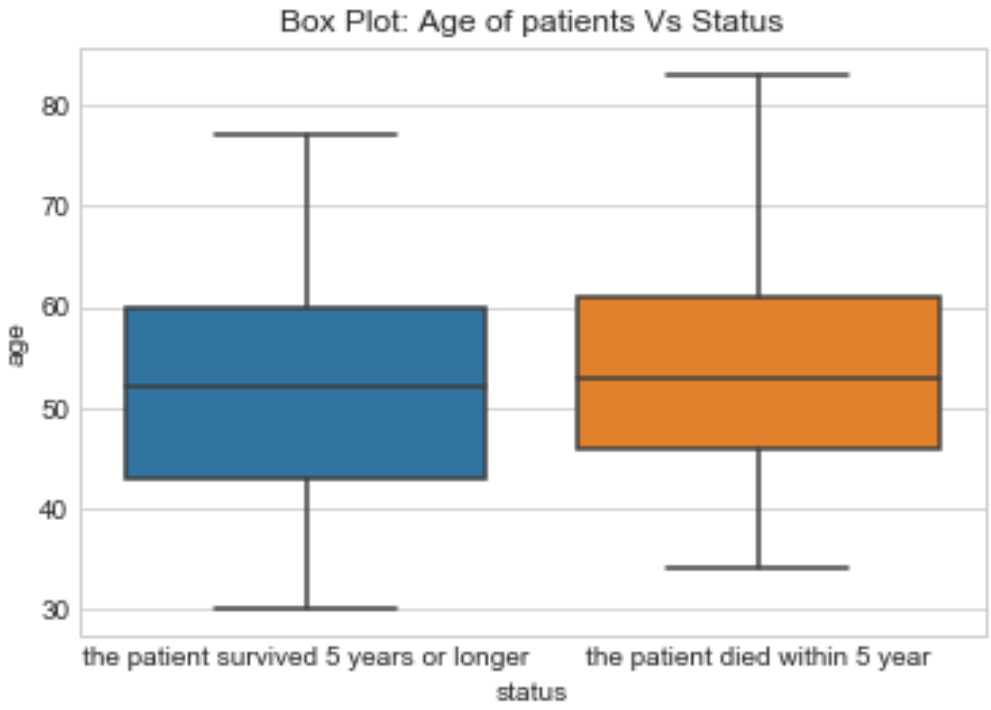


Observation:

CDF of age of patients vs survived and vs died plots are closely related hence "AGE" alone cannot be a determining factor for survival.

CDF of number of positive axillary nodes detected vs survived and vs died plots are closely related hence "NODES detected" alone cannot be a determining factor for survival.

```
#Univariant Analysis : Box plots

sns.boxplot(x='status',y='age', data=haberman)
plt.title("Box Plot: Age of patients Vs Status")
plt.show()
```

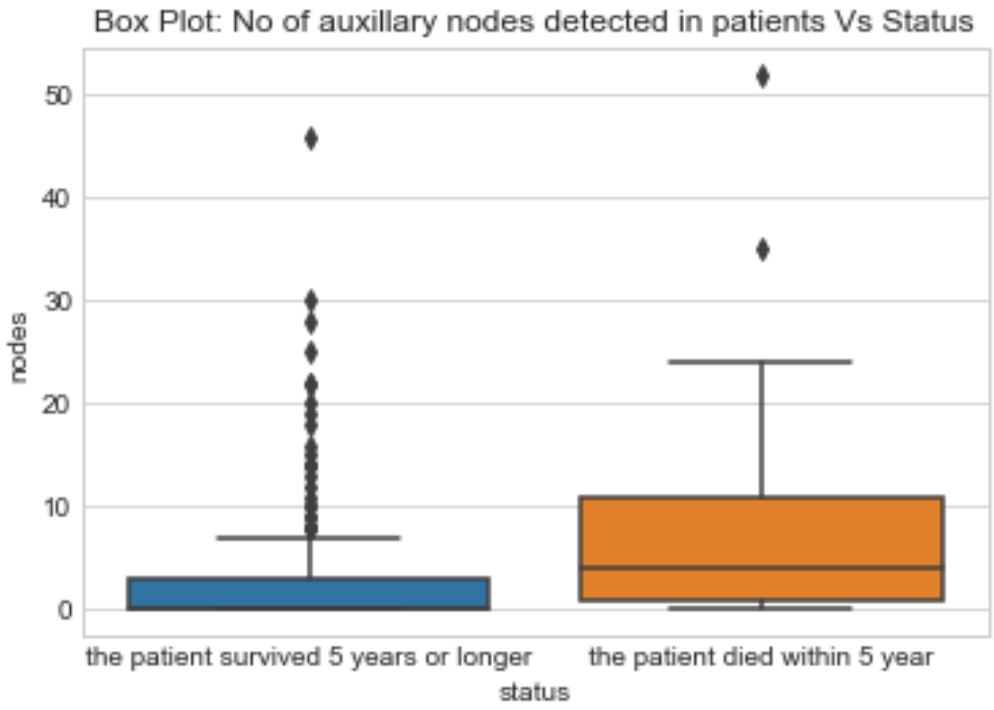


Observation:

Approximate observations: Median - 52; age range - 42 to 62

Two box plots are closely ploted accross ages hence age alone cannot be sufficient to understand the survival rate

```
sns.boxplot(x='status',y='nodes', data=haberman)
plt.title("Box Plot: No of auxillary nodes detected in patients Vs Status")
plt.show()
```

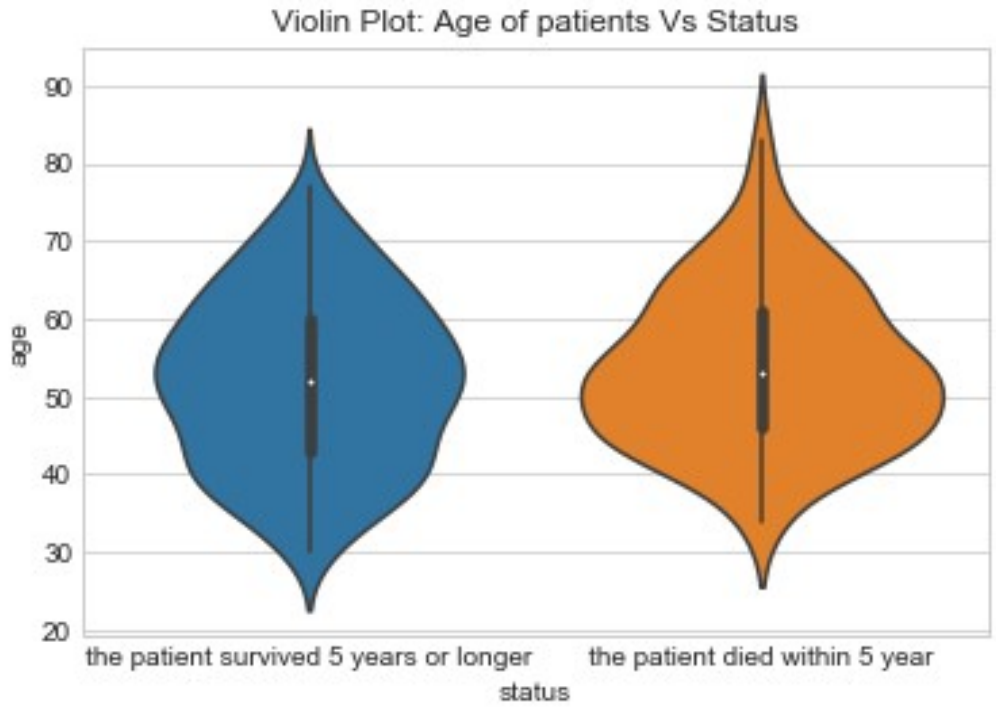


Observation:

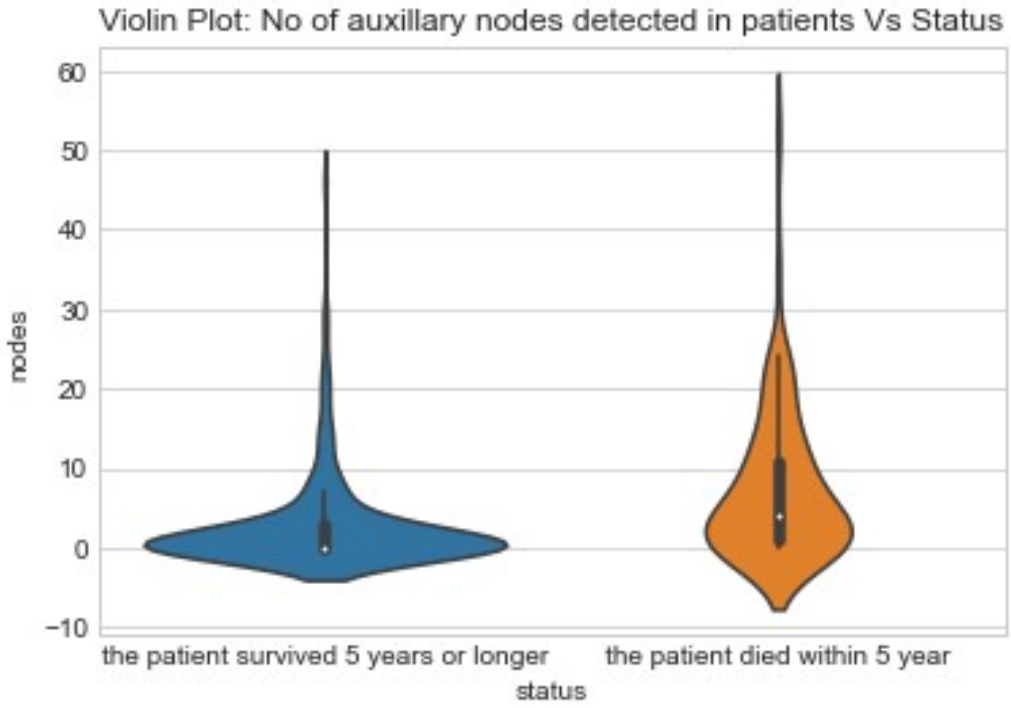
Patients survived 5 years or longer are most likely the ones with 0 nodes detected when compared to more nodes detected. Hence lesser the nodes detected more the chances of survival.

But still nodes detected alone cannot be sufficient to understand the survival rate as two box plots are closely ploted across y-axis

```
#Uni-Variant Analysis: ViolinPlots
sns.violinplot(x="status", y="age", data=haberman, size=8)
plt.title("Violin Plot: Age of patients Vs Status")
plt.show()
```



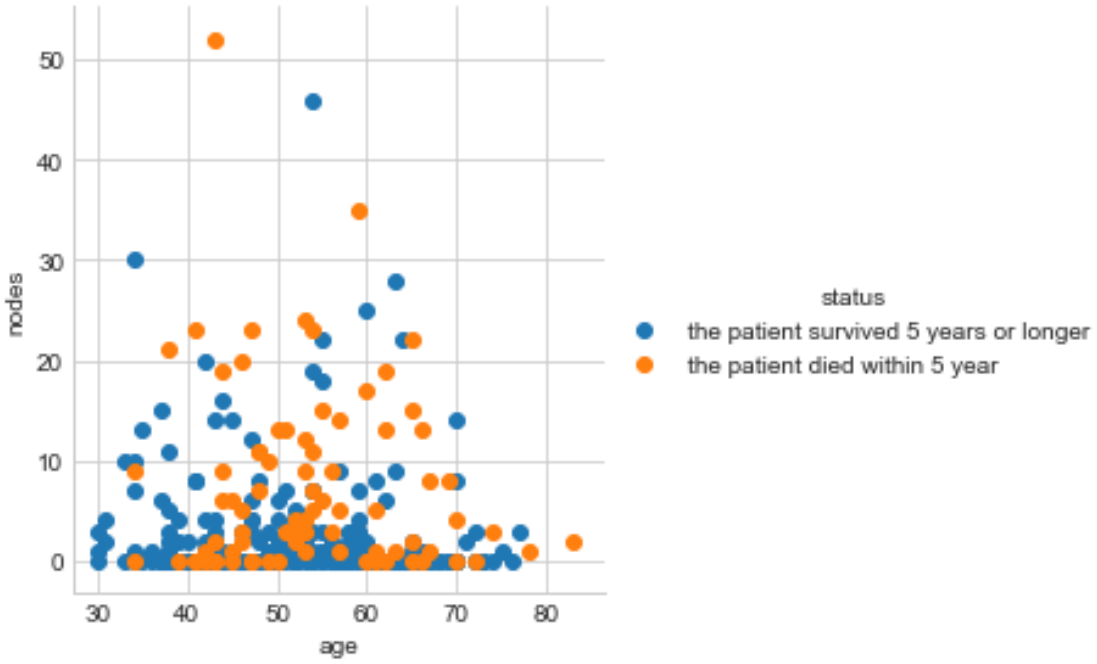

```
sns.violinplot(x="status", y="nodes", data=haberman, size=8)
plt.title("Violin Plot: No of auxillary nodes detected in patients Vs
Status")
plt.show()
```



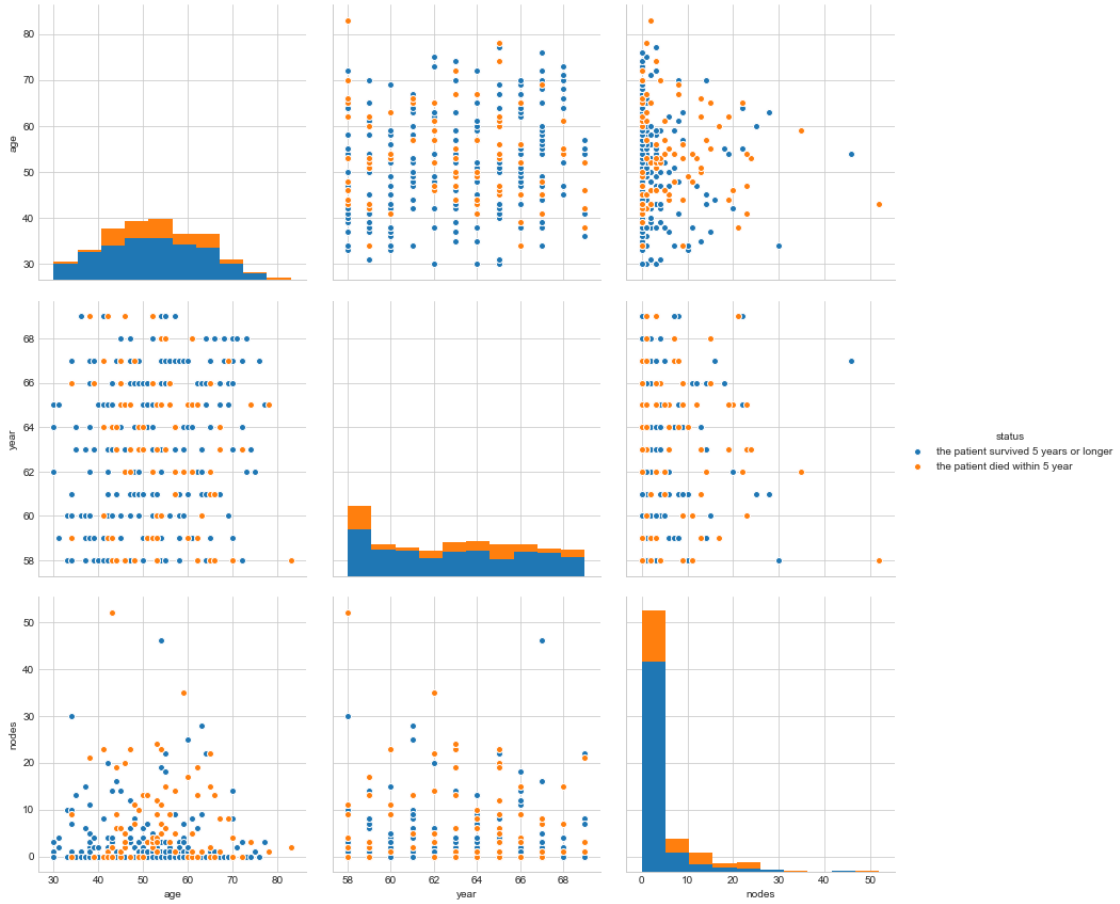
Observation:

Two violin plots are closely ploted across y-axis hence age and nodes detected alone cannot be sufficient to understand the survival rate. Need to do Bi-Variant analysis

```
#Bi-Variant Analysis: 2-D Scatter Plot

sns.FacetGrid(haberman, hue="status", size=4) \
   .map(plt.scatter, "age", "nodes") \
   .add_legend()

plt.show()
```




```
#Bi-Variant Analysis: 3-D Scatter Plot : Pair Plots

plt.close();
sns.set_style("whitegrid");
sns.pairplot(haberman, hue="status", size=4);
plt.show()
```



Final Observations:

30<age<40 : High survival of 5 years or longer irrespective of nodes detected

40<age<50 : if nodes detected =0 or less than 10 higher chances to die with in 5 years of surgery

50<age<60 : if nodes dectected =0 then high survival of 5 years or longer irrespective of nodes detected

60<age<70 : lesser survival i.e., higher chances to die with in 5 years of surgery

All these observations are made out of nodes vs age scatter plot.