exploratory_data_analysis_exercise

September 24, 2021

0.1 Exercise:

```
Download Haberman Cancer Survival dataset from Kaggle. You may have to create a Kaggle account
Perform a similar alanlaysis as above on this dataset with the following sections:

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

[1]:

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

#Load Haberman Cancer Survival dataset
hcs = pd.read_csv("haberman.csv")
```

[2]:

```
#How Many data-points and features
print(hcs.shape)
```

```
(306, 4)
```

we have 306 rows and 4 columns

[3]:

```
#number of classes
print(hcs.columns)
```

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

[4]:

```
hcs.head()
```

[4]:

```
   age  year  nodes  status
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]: hcs.tail()
```

```
[5]:      age  year  nodes  status
     301   75    62      1       1
     302   76    67      0       1
     303   77    65      3       1
     304   78    65      1       2
     305   83    58      2       2
```

```
[6]: #data-points per class
     hcs['year'].value_counts()
```

```
[6]: 58    36
     64    31
     63    30
     66    28
     65    28
     60    28
     59    27
     61    26
     67    25
     62    23
     68    13
     69    11
     Name: year, dtype: int64
```

```
[7]: hcs['age'].value_counts()
```

```
[7]: 52    14
     54    13
     50    12
     47    11
     53    11
     43    11
     57    11
     55    10
     65    10
     49    10
     38    10
     41    10
     61     9
     45     9
     42     9
     63     8
     59     8
     62     7
     44     7
```

```
58     7
56     7
46     7
70     7
34     7
48     7
37     6
67     6
60     6
51     6
39     6
66     5
64     5
72     4
69     4
40     3
30     3
68     2
73     2
74     2
36     2
35     2
33     2
31     2
78     1
71     1
75     1
76     1
77     1
83     1
Name: age, dtype: int64
```

[8]:

```
hcs['nodes'].value_counts()
```

[8]:

```
0     136
1      41
2      20
3      20
4      13
6       7
7       7
8       7
5       6
9       6
13      5
14      4
11      4
```

```
10      3
15      3
19      3
22      3
23      3
12      2
20      2
46      1
16      1
17      1
18      1
21      1
24      1
25      1
28      1
30      1
35      1
52      1
Name: nodes, dtype: int64
```

```
[9]: hcs['status'].value_counts()
```

```
[9]: 1    225
     2     81
     Name: status, dtype: int64
```

Analysis of survival of patients who had undergone surgery 225 patients survived more than 5 years and 81 patients survived less than 5 years

1 Bi-variate analysis

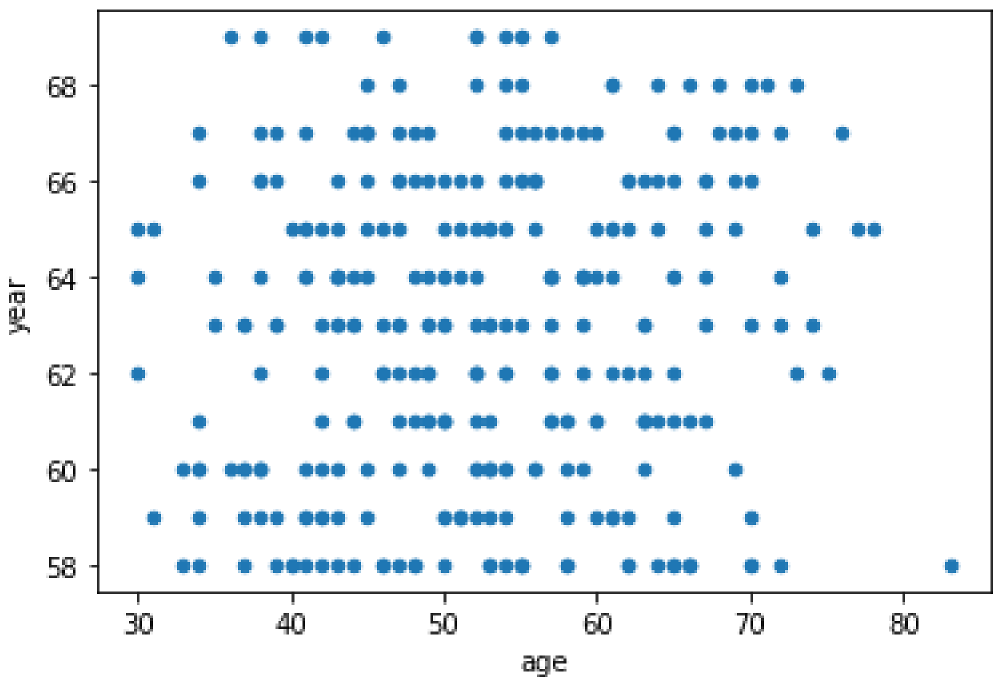
1.1 2-D Scatter Plot

```
[10]: print(hcs.columns)
```

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

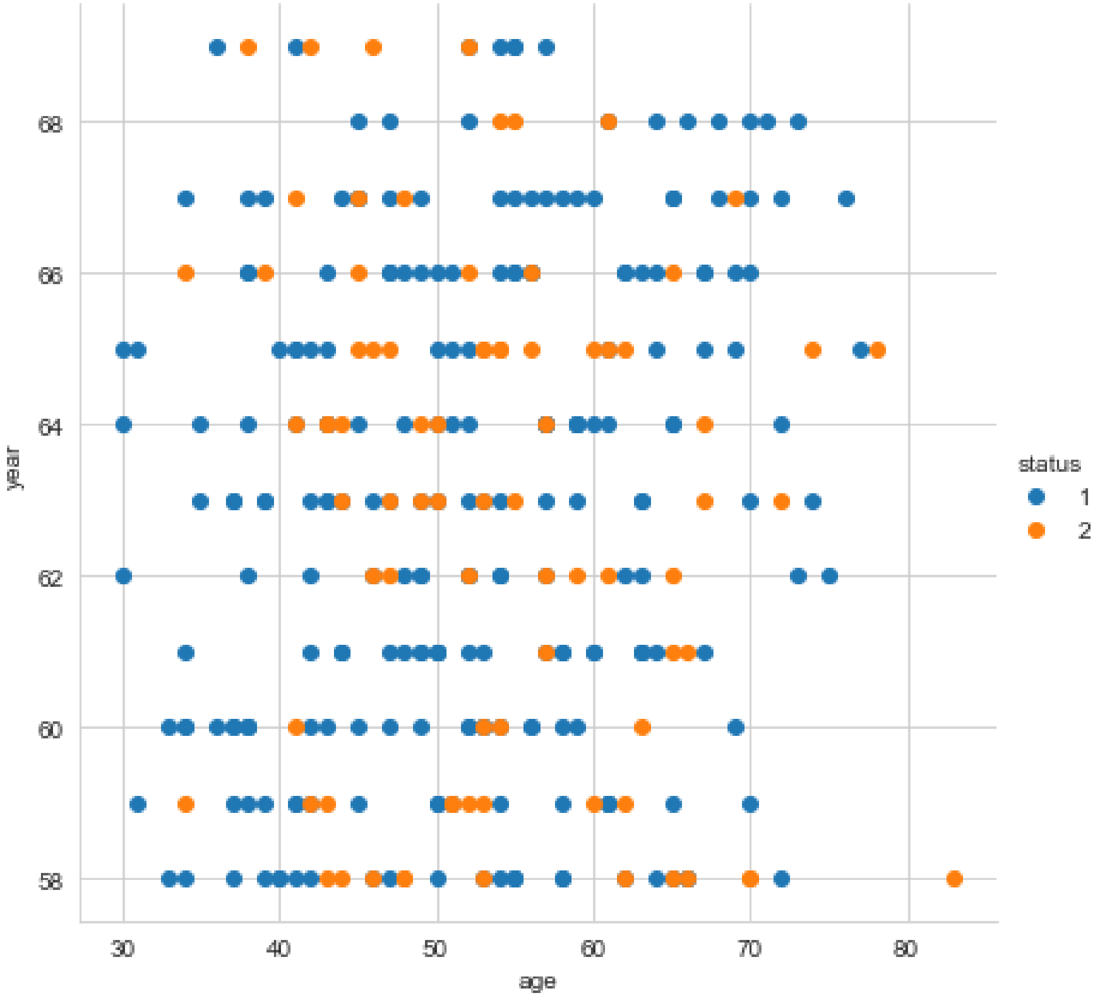
```
[11]: hcs.plot(kind="scatter", x='age', y='year')
```

```
[11]: <AxesSubplot:xlabel='age', ylabel='year'>
```



[12]:
```
#2-D scatter plot with color code of survival list
sns.set_style("whitegrid");
sns.FacetGrid(hcs, hue="status", size=6).map(plt.scatter,"age","year").
 ↪add_legend();
plt.show()
```

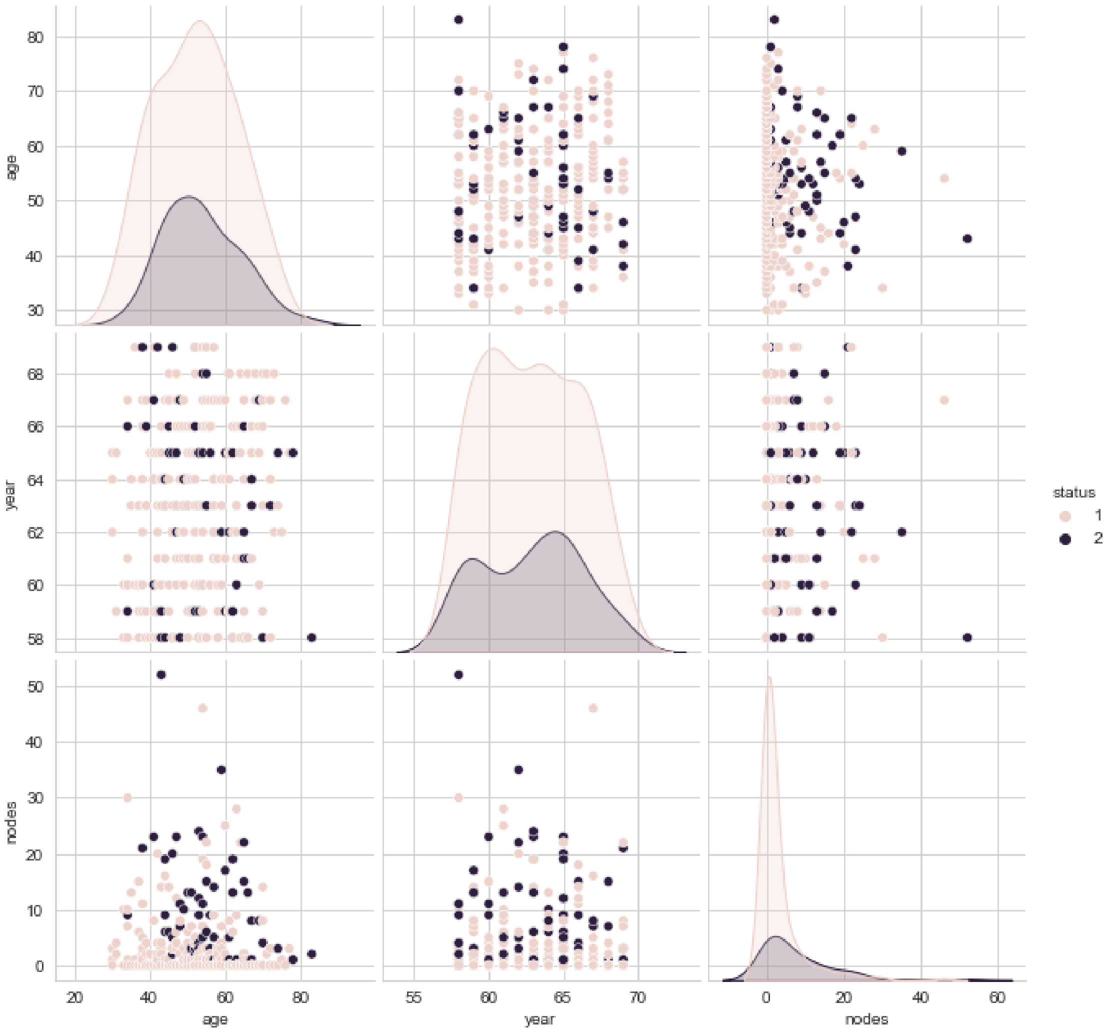
```
C:\Users\Shadrach\anaconda3\lib\site-packages\seaborn\axisgrid.py:316:
UserWarning: The `size` parameter has been renamed to `height`; please update
your code.
  warnings.warn(msg, UserWarning)
```



1.2 Pair-plot

[13]:
```
plt.close();
sns.set_style("whitegrid");
sns.pairplot(hcs, hue="status", size=3).add_legend;
plt.show()
```

```
C:\Users\Shadrach\anaconda3\lib\site-packages\seaborn\axisgrid.py:1912:
UserWarning: The `size` parameter has been renamed to `height`; please update
your code.
  warnings.warn(msg, UserWarning)
```



1.2.1 Observations:

Survival status is random spread but based on analysis

- There is high survival between age group 30 to 40
- pateient have nodes between 0 and 1 has high survival and patients having node range 25 has very less survival

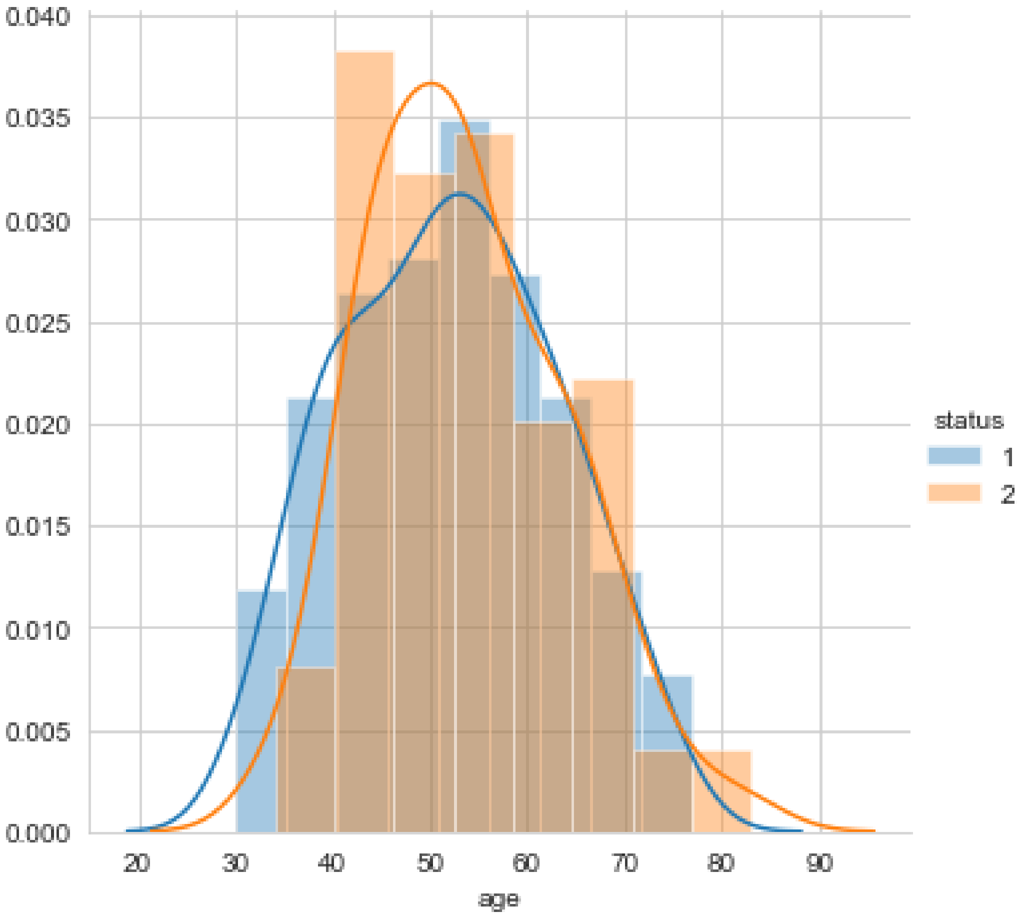
2 Univaraite analysis

2.0.1 PDF

[14]:
```
hcs_surv=hcs.loc[hcs["status"] == 1];
hcs_canc=hcs.loc[hcs["status"] == 2];

sns.FacetGrid(hcs, hue='status',height=5).map(sns.distplot,'age').add_legend();
```

```
plt.show()
```

```
C:\Users\Shadrach\anaconda3\lib\site-packages\seaborn\distributions.py:2551:
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
C:\Users\Shadrach\anaconda3\lib\site-packages\seaborn\distributions.py:2551:
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```

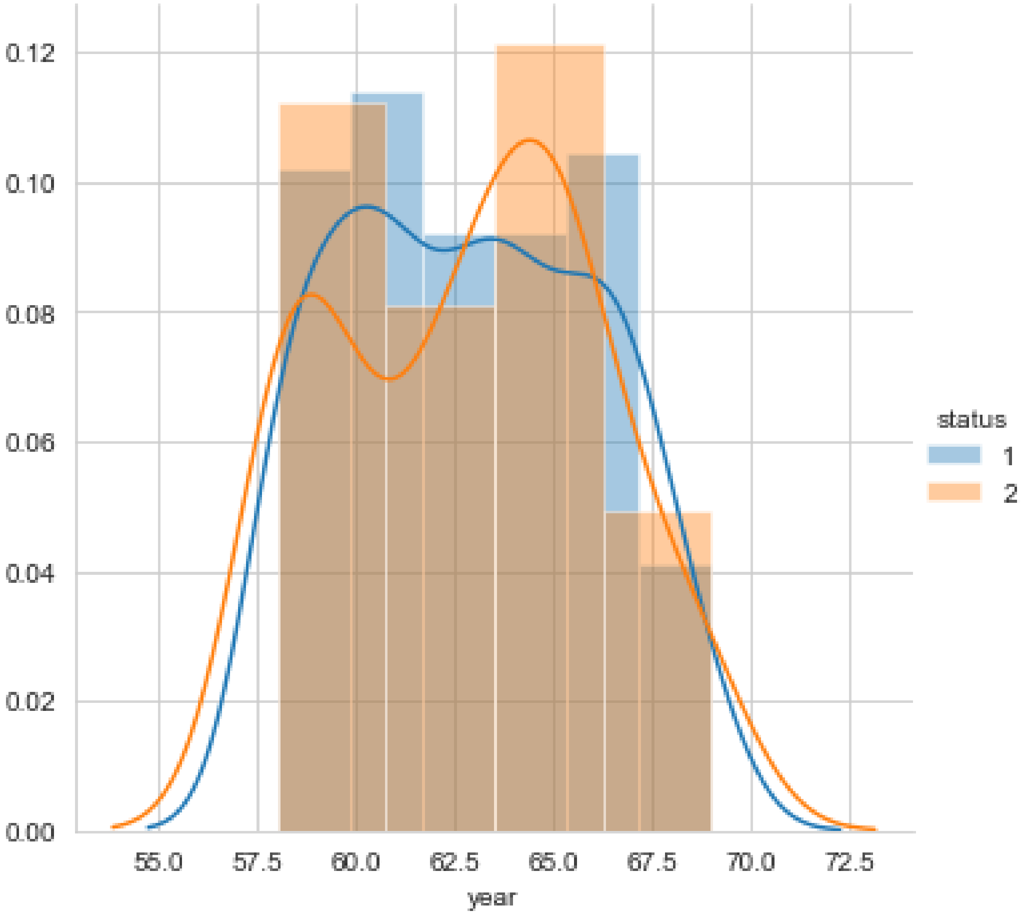


[15]:

```
sns.FacetGrid(hcs, hue='status',height=5).map(sns.distplot,'year').add_legend();
plt.show()
```

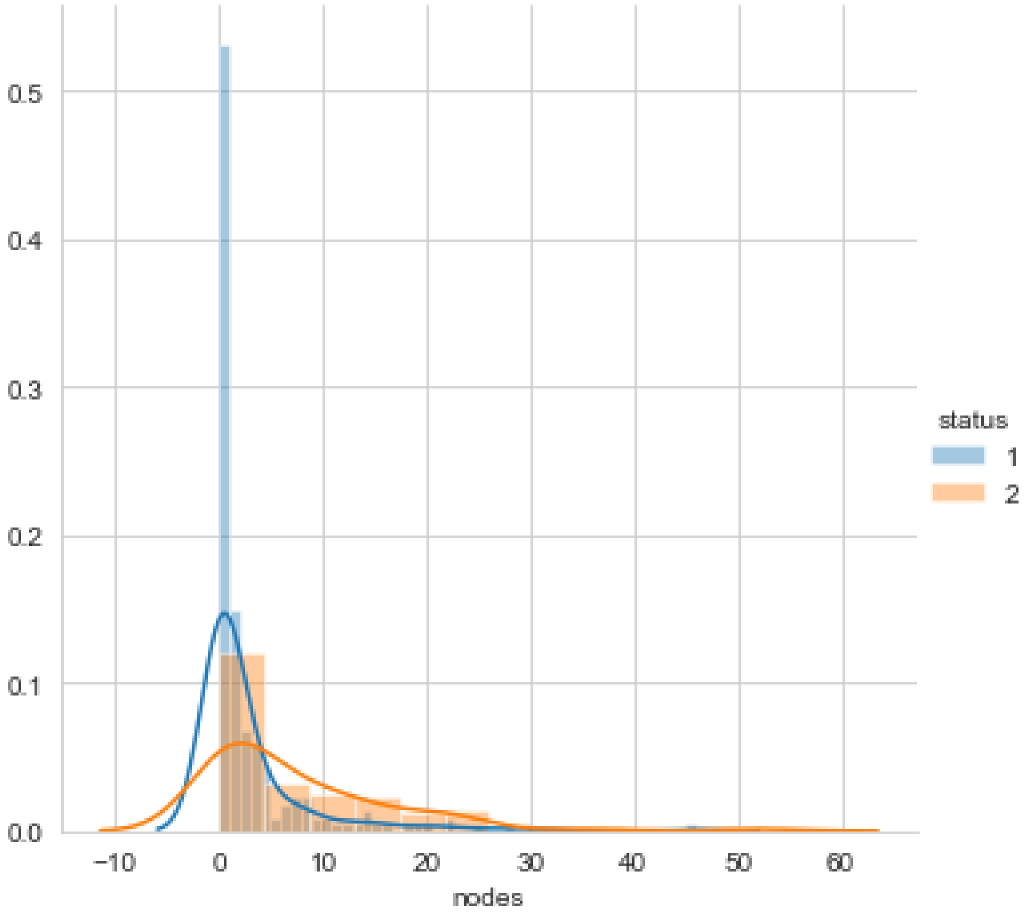
```
C:\Users\Shadrach\anaconda3\lib\site-packages\seaborn\distributions.py:2551:
```

```
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
C:\Users\Shadrach\anaconda3\lib\site-packages\seaborn\distributions.py:2551:
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```



[16]:
```
sns.FacetGrid(hcs, hue='status',height=5).map(sns.distplot,'nodes').
 ↪add_legend();
plt.show()
```

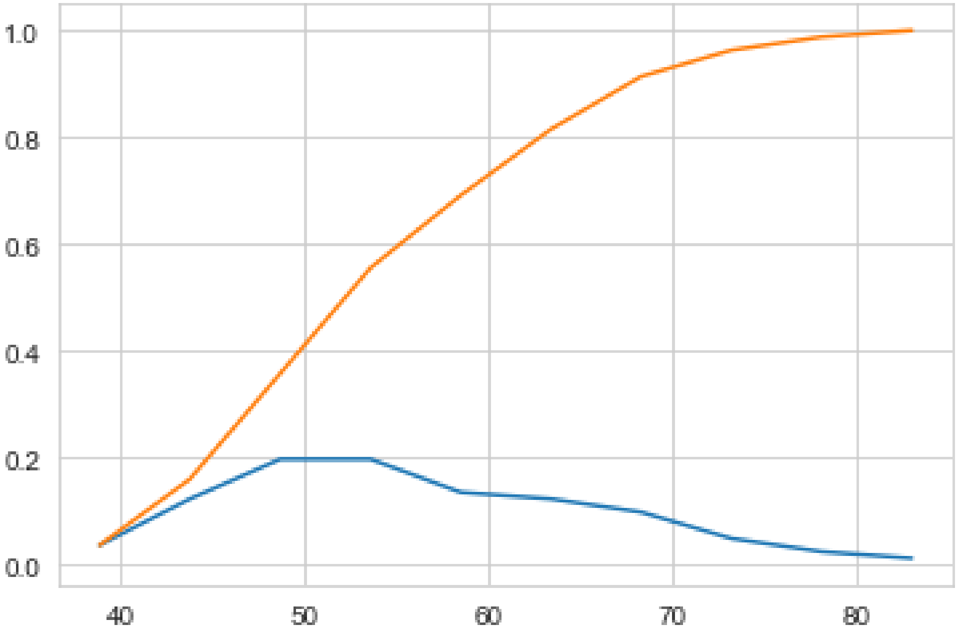
```
C:\Users\Shadrach\anaconda3\lib\site-packages\seaborn\distributions.py:2551:
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
C:\Users\Shadrach\anaconda3\lib\site-packages\seaborn\distributions.py:2551:
FutureWarning: `distplot` is a deprecated function and will be removed in a
future version. Please adapt your code to use either `displot` (a figure-level
function with similar flexibility) or `histplot` (an axes-level function for
histograms).
  warnings.warn(msg, FutureWarning)
```



[17]:
```python
counts, bin_edges = np.histogram(hcs_canc['age'], bins=10, density= True);

pdf = counts/(sum(counts));
print(pdf);
print(bin_edges);
cdf=np.cumsum(pdf);

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

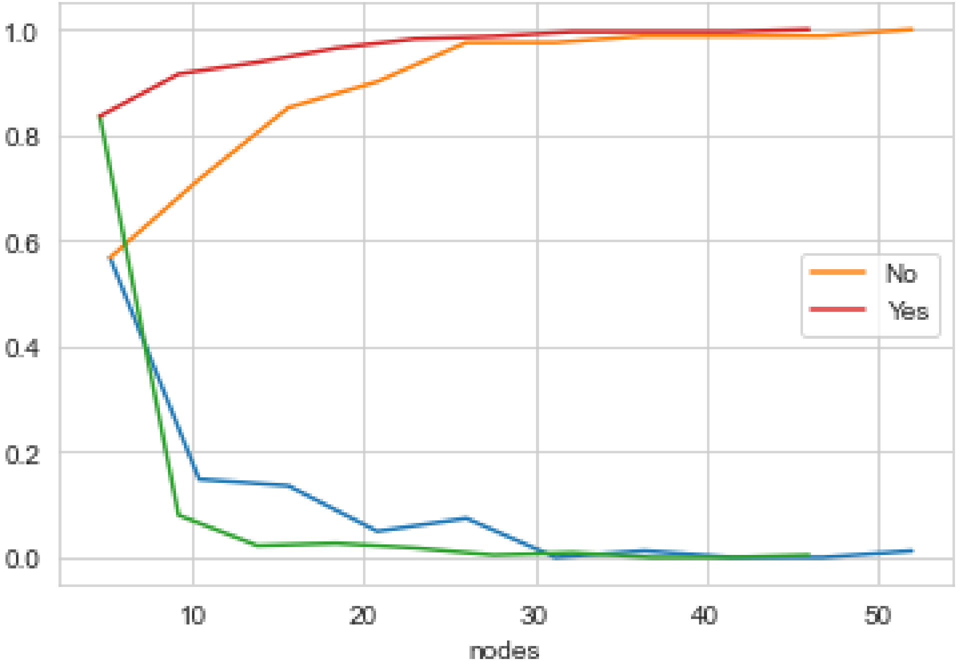
```
[0.03703704 0.12345679 0.19753086 0.19753086 0.13580247 0.12345679
 0.09876543 0.04938272 0.02469136 0.01234568]
[34.  38.9 43.8 48.7 53.6 58.5 63.4 68.3 73.2 78.1 83. ]
```

```
[17]: [<matplotlib.lines.Line2D at 0x232cefebd90>]
```



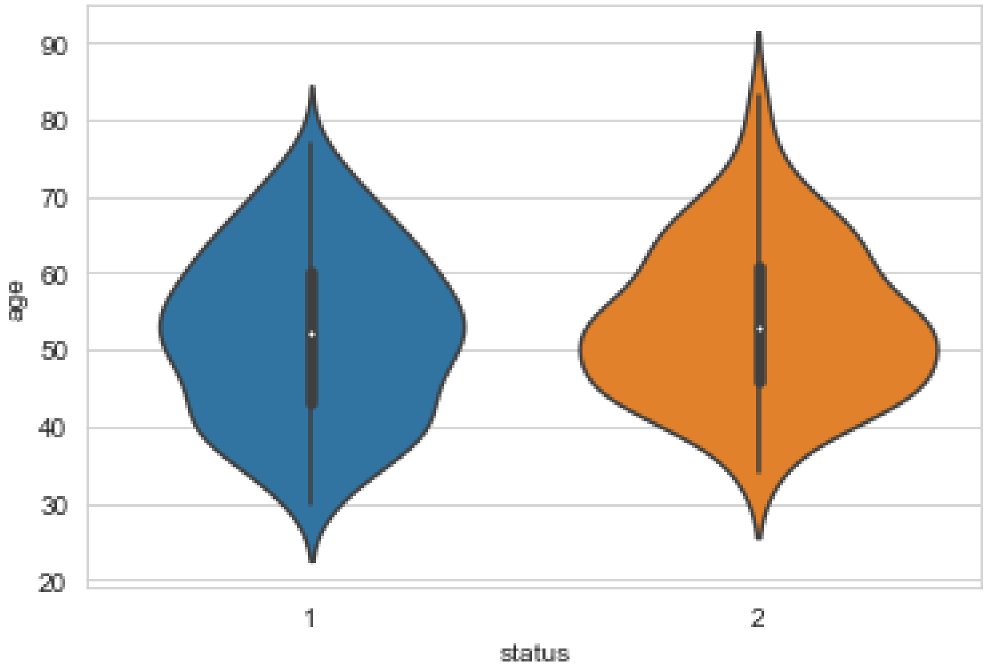
```
[18]: #cancer
counts1, bin_edges1 = np.histogram(hcs_canc['nodes'], bins=10, density= True);

pdf1 = counts1/(sum(counts1));
print(pdf1);
print(bin_edges1);
cdf1=np.cumsum(pdf1);

sns.set_style("whitegrid");
plt.plot(bin_edges1[1:],pdf1);
plt.plot(bin_edges1[1:], cdf1, label = 'No');
plt.xlabel('nodes')

#survival
counts2, bin_edges2 = np.histogram(hcs_surv['nodes'], bins=10, density= True);

pdf2 = counts2/(sum(counts2));
print(pdf2);
print(bin_edges2);
cdf2=np.cumsum(pdf2);

plt.plot(bin_edges2[1:],pdf2);
plt.plot(bin_edges2[1:], cdf2, label='Yes');
plt.xlabel('nodes');

plt.legend();
plt.show();
```

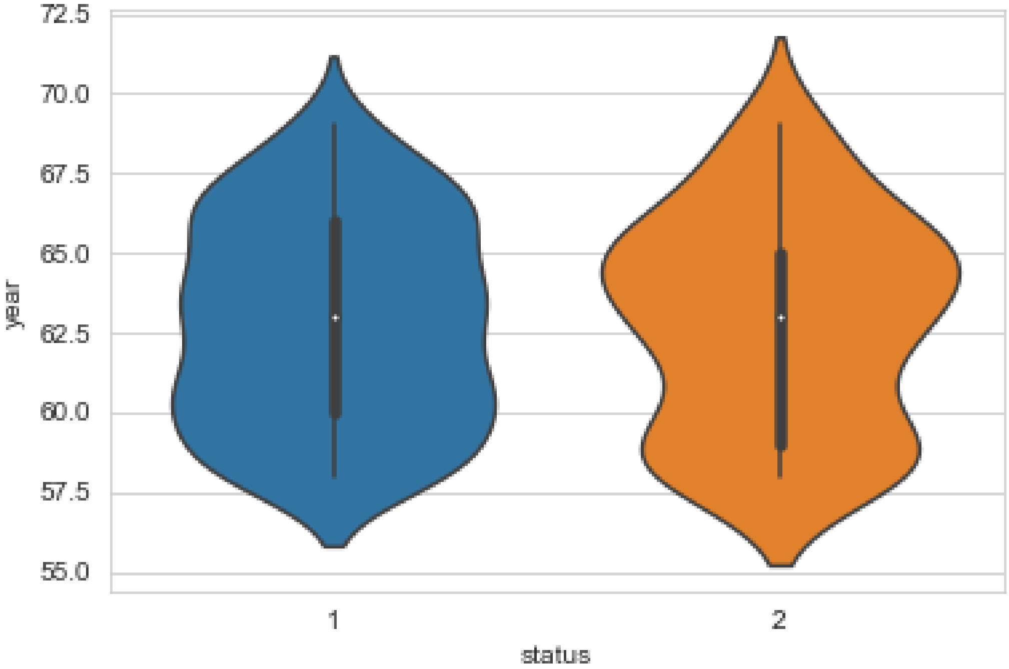
```
[0.56790123 0.14814815 0.13580247 0.04938272 0.07407407 0.
 0.01234568 0.         0.         0.01234568]
[ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
[0.83555556 0.08       0.02222222 0.02666667 0.01777778 0.00444444
 0.00888889 0.         0.         0.00444444]
[ 0.   4.6  9.2 13.8 18.4 23.  27.6 32.2 36.8 41.4 46. ]
```

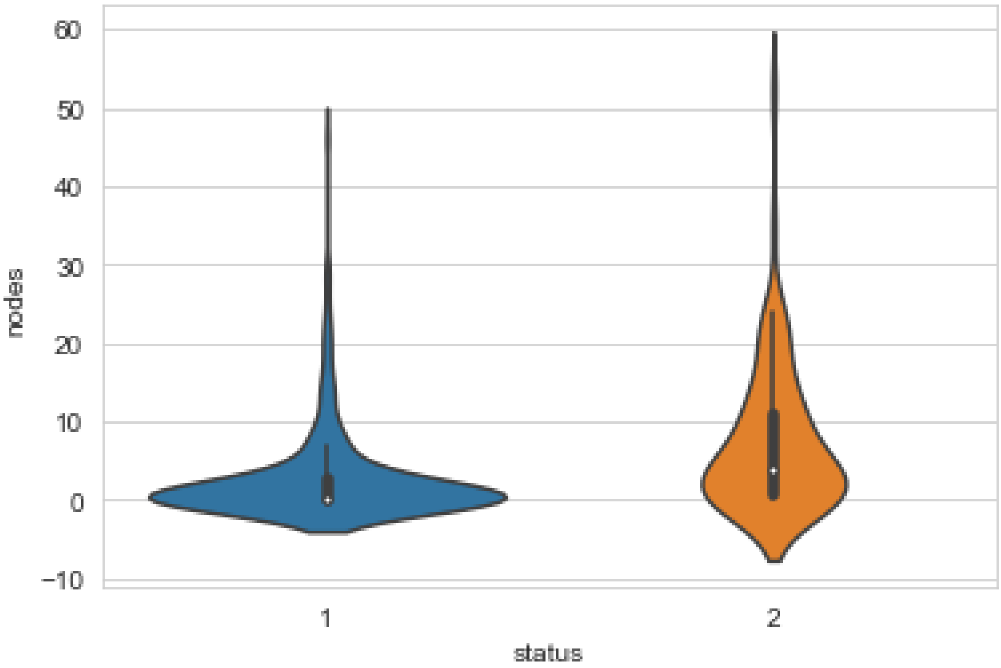


2.1 Voilin Plot

```
[19]: sns.violinplot(x="status", y="age", data=hcs, size=8)
plt.show()
sns.violinplot(x="status", y="year", data=hcs, size=8)
plt.show()
sns.violinplot(x="status", y="nodes", data=hcs, size=8)
plt.show()
```


90
80
70
60
50
40
30
20
age
1
2
status

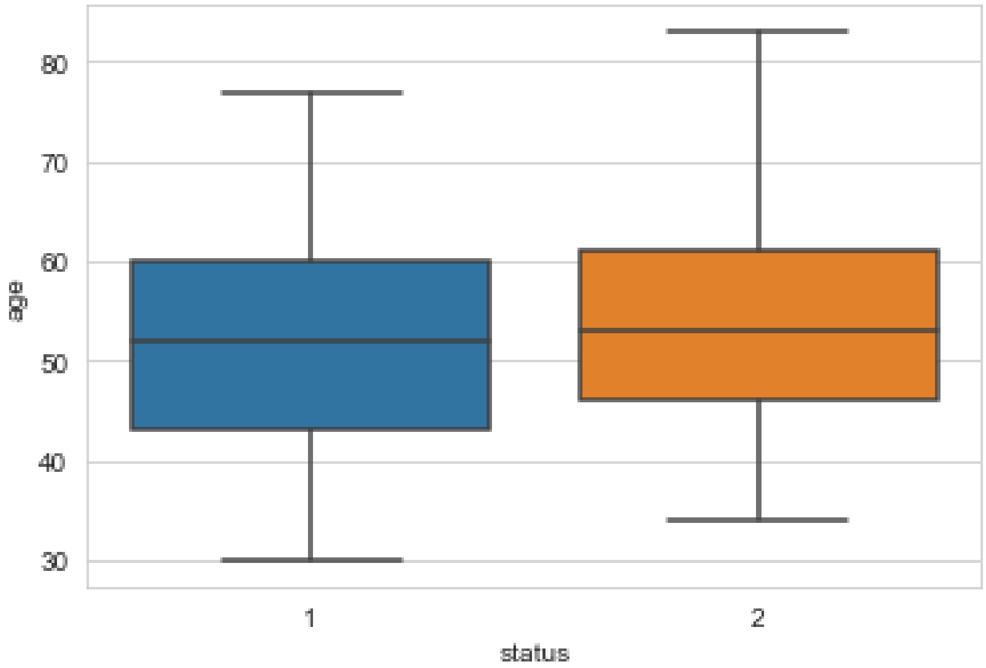

72.5
70.0
67.5
65.0
62.5
60.0
57.5
55.0
year
1
2
status

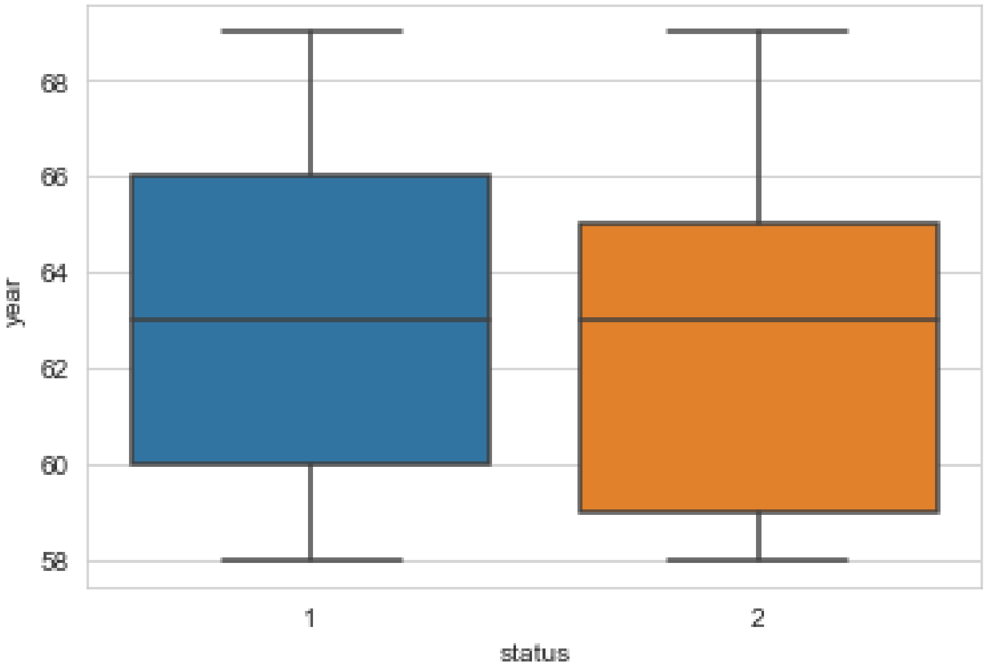


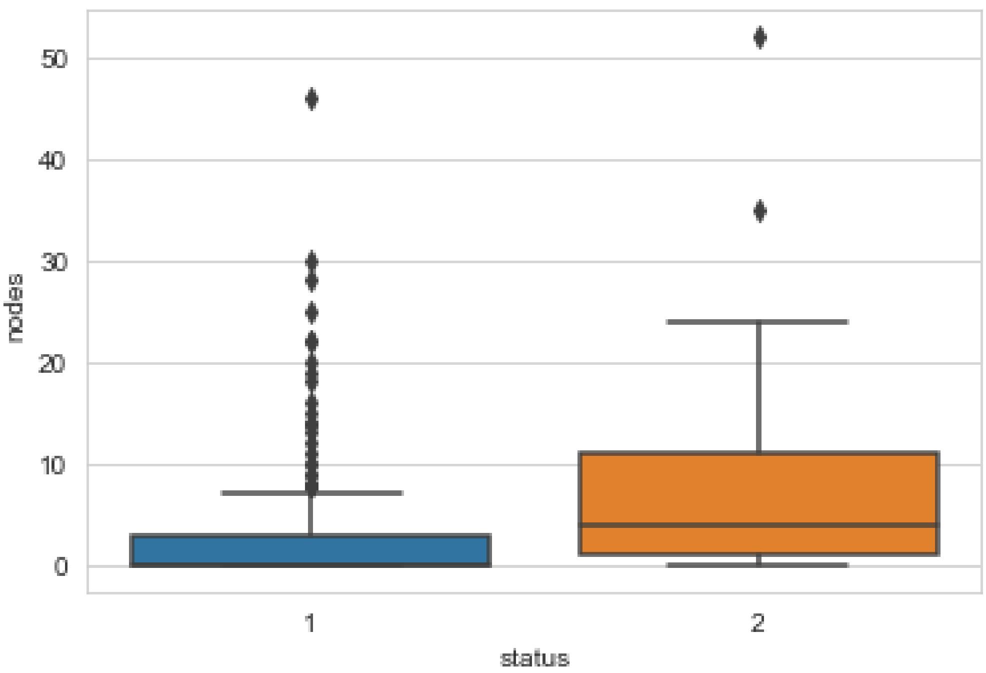
2.2 BOX PLOT

[20]:

```
sns.boxplot(x="status", y="age", data=hcs)
plt.show()
sns.boxplot(x="status", y="year", data=hcs)
plt.show()
sns.boxplot(x="status", y="nodes", data=hcs)
plt.show()
```


80
70
60
50
40
30
age
1
2
status


68
66
64
62
60
58
year
1
2
status



2.2.1 Observations:

- we have 306 rows and 4 columns
- data set has following columns ('age', 'year', 'nodes', 'status')
- Status is a feature which represent 1 -> survived, 2 -> no
- From the dataset we understood that 221 survived and 81 didn't

###Post univariate and bivariate analysis

- There is high survival between/less age group 30 to 40
- patient having nodes between 0 and 1 has high survival and patients having node range >=25 has very less survival
- Between 1960 and 1965 there were more unsuccessful operations.

###Conclusion

- These aren't the deciding factors but however there were non-survival cases between age group 30 to 40 but people less than 35 years have more chance of survival
- The objective of classifying the survival status of a new patient based on the given features is a difficult task as the data is imbalanced.