EDA_of_HabermanDataset

July 25, 2020

1 Exploratory Data Analysis of Haberman Dataset

The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

1.1 Objective:

- To analyse the Haberman Dataset and record observation.
- To find the survival status of a patient whose features value maybe given and whose data don't belong to the dataset.

[ ]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as mtplot
import numpy as np

#Reading Haberman.csv using pandas
haberman = pd.read_csv('haberman.csv')

#Knowing about datapoints and attributes
print('Datapoints = ',haberman.shape[0])
print('Attributes = ',haberman.shape[1] )
```

```
Datapoints =  306
Attributes =  4
```

We can say from the above output,

- Haberman dataset has 306 rows and 4 columns.
- Rows are called datapoints and columns are called Attributes. All attributes except the last attribute are called features/variables. Features do not include the last column because they are independent variables. The last column is called Class Label or Response Label because it is the output value or the dependant variable.
- So Haberman dataset has 3 independent variable and 1 dependent variable.

[ ]:
```
#Listing the Attributes and class label of Haberman dataset
print(haberman.columns)
```

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

We can say from the above output, * Features of Haberman dataset are: 1. age 2. year 3. nodes * **status** is the Class Label/Dependant Variable/ Output Value/Response Label

2 Significance of the attributes of Haberman dataset.

1. age = Age of the patient at the time of operation
2. year = Year at which the operation has taken place (year - 1900)
3. nodes = Number of Positive axil nodes detected during operation
4. status = survival status of the patient
   * If (status = 1): Patient survived 5 years or longer
   * If (status = 2): Patient died within 5 years.

[ ]:
```
# Calculating total number of datapoints each class label is having.
haberman['status'].value_counts()
```

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

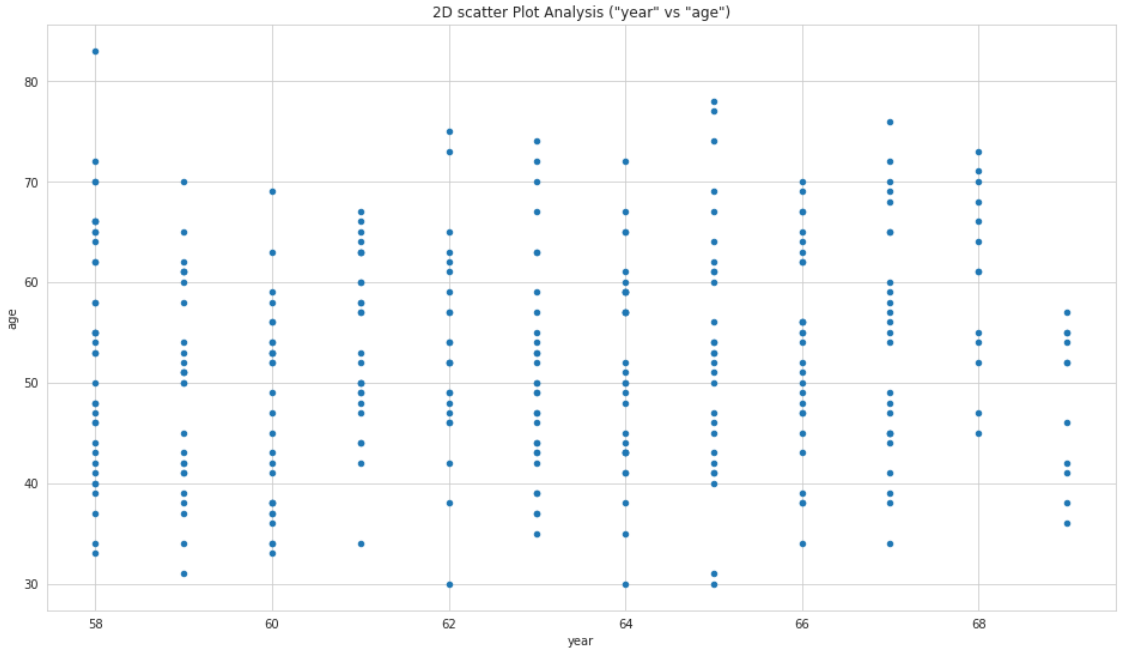
Observations

- 225 datapoints have status = Survived 5 years or longer.
- 81 datapoints have status = Died within 5 years.
- This is a **Balanced Dataset** because the difference between two class labels are not huge.
- If the difference was **huge**, the dataset would have been called as an **Imbalanced Dataset**.
- Also it reveals patients who survived more than 5 years are greater than patients who died with in 5 years after operation.

3 Bivariate Analysis

3.1 2D Scatter Plot

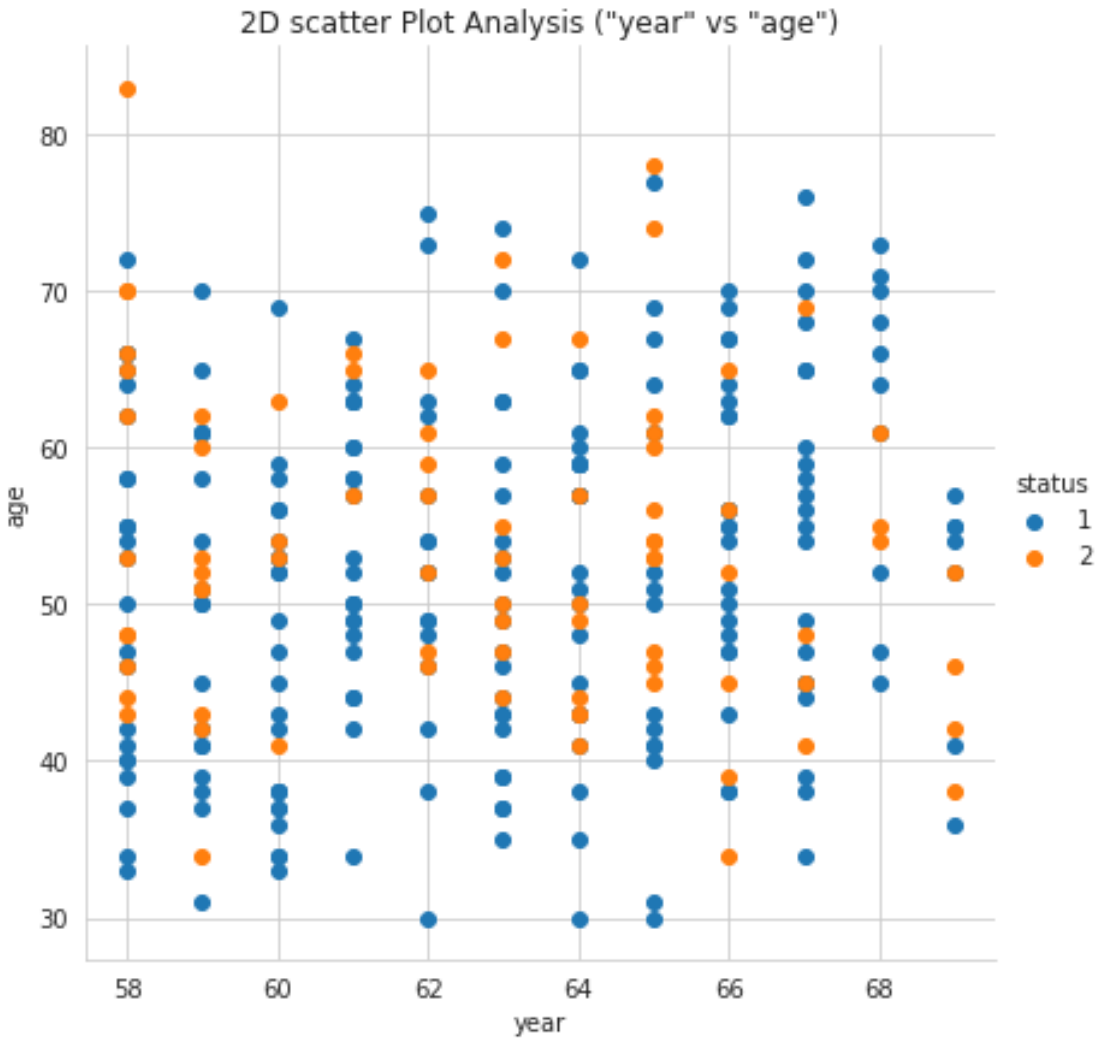
[ ]:
```
# Plotting 'year' vs 'age' graph to analyse haberman dataset
sns.set_style('whitegrid')
haberman.plot(kind='scatter', x='year', y='age')
mtplot.title('2D scatter Plot Analysis ("year" vs "age")')
mtplot.show()
```



Observation

1. Women of age group 30 yrs to 80+ yrs are found to have breast cancer.

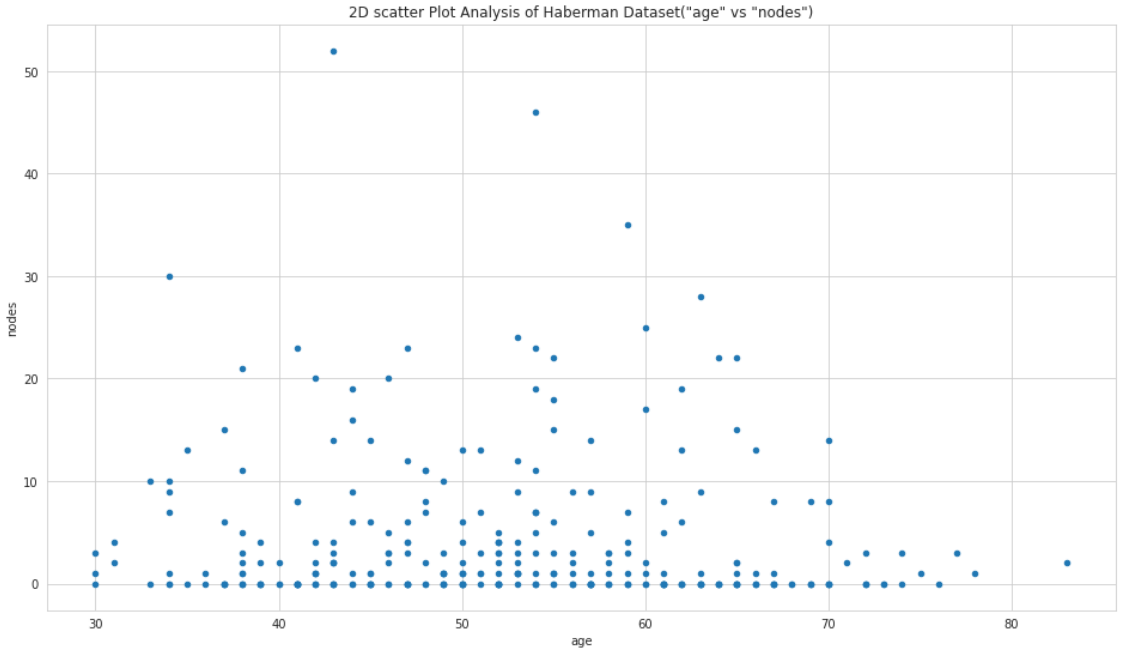
[ ]:
```
# Plotting 'year' vs 'age' graph to analyse haberman dataset but with colorful␣
 ↪plot points
sns.set_style('whitegrid')
sns.FacetGrid(haberman, hue = 'status', height = 6).map(mtplot.
 ↪scatter,'year','age').add_legend()
mtplot.title('2D scatter Plot Analysis ("year" vs "age")')
mtplot.show()
```



Observation

- No straight line can be drawn to differentaite between survived and death cases.
- More death cases are observed among women of age 41yrs to 68yrs.
- Fewer death cases are obsderved among women of age 70+ yrs and 30yrs. to 40yrs.

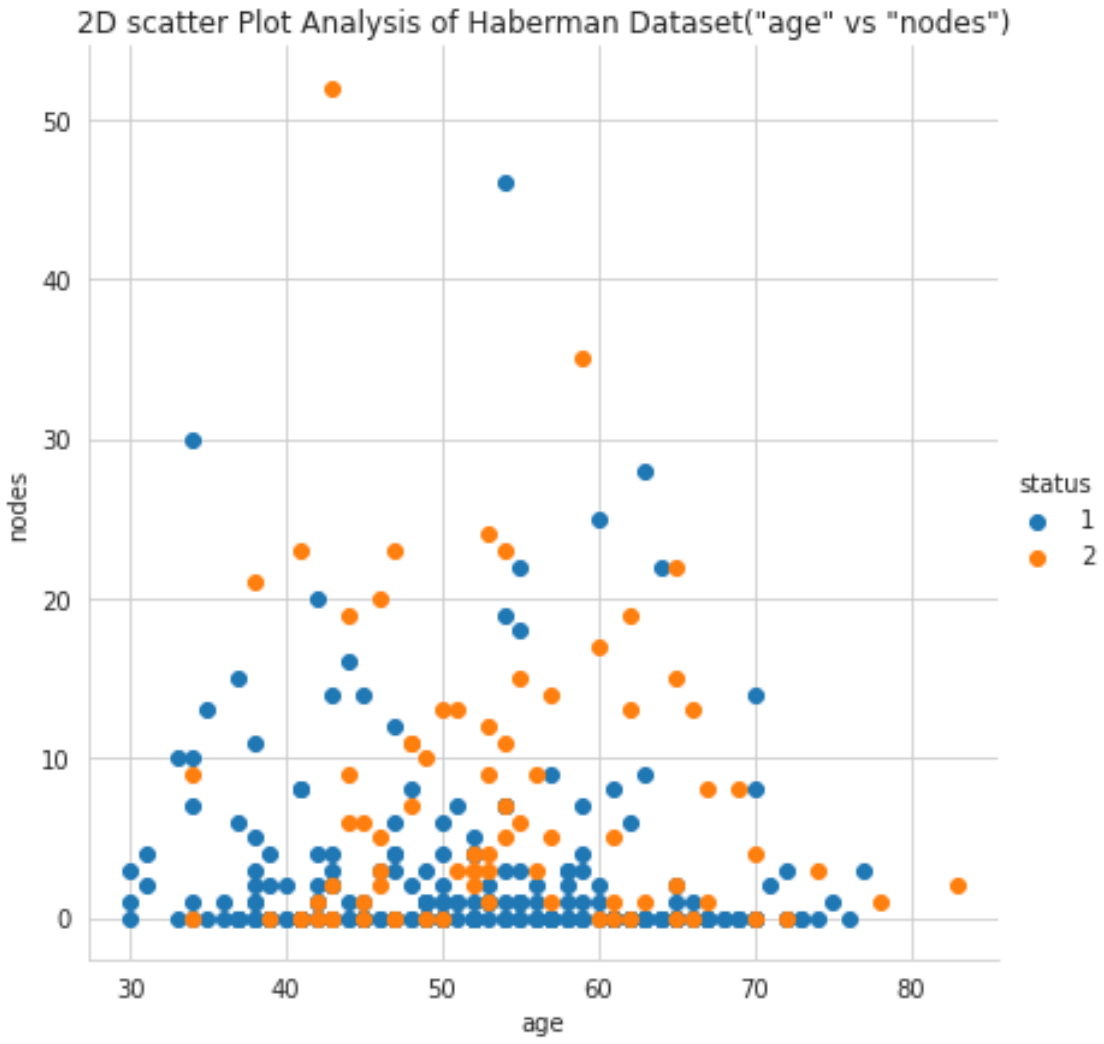
[137]:
```
# Plotting 'age' vs 'nodes' graph to analyse haberman dataset to reveal more␣
↪information.
sns.set_style('whitegrid')
haberman.plot(kind = 'scatter', x = 'age', y = 'nodes')
mtplot.title('2D scatter Plot Analysis of Haberman Dataset("age" vs "nodes")')
mtplot.show()
```



Observations

- Highest number of Axillary node detected is greater than 50
- Most patients have axillary nodes ranging betwee 0 to 10.
- Most women have positive axillary nodes ranging from 0 to 6 in age group 30yrs to 78yrs having greater density in age group of 48yrs to 58yrs.
- Moderate number of women have positive axillary nodes ranging between 6 to 15.
- Few had greater than 15 nodes.

[136]:
```
sns.set_style('whitegrid')
sns.FacetGrid(haberman, hue = 'status', height = 6).map(mtplot.
 ↪scatter,'age','nodes').add_legend()
mtplot.title('2D scatter Plot Analysis of Haberman Dataset("age" vs "nodes")')
mtplot.show()
```



Observation

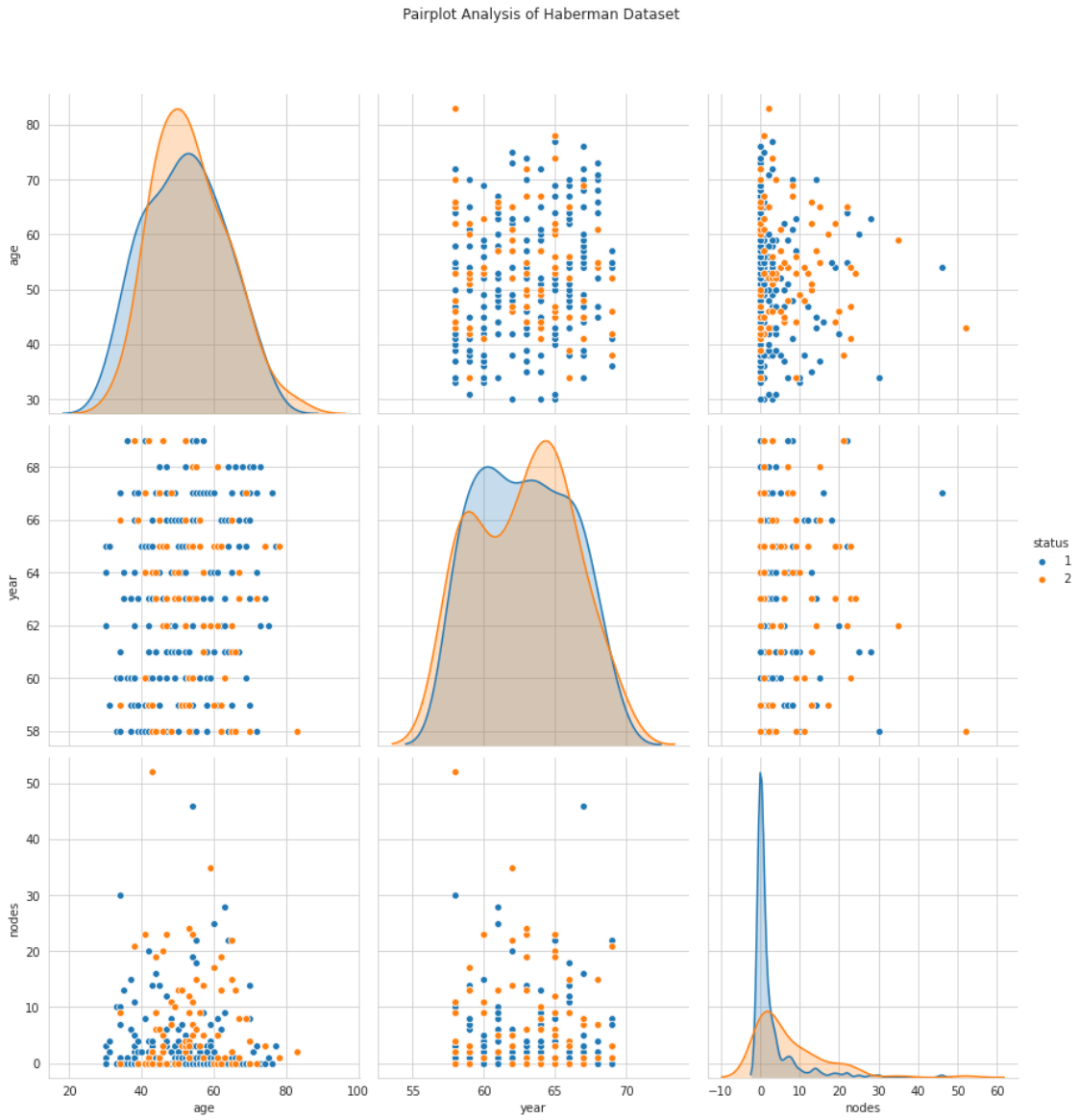
- No straight line can be drawn to differentiate between the class labels of the dataset.
- But more deaths are 40 to 70 yrs with positive axillary nodes ranging between 0 to 25, most having nodes range of 0 to 9.
- Age group 48yrs to 66yrs. have mostly 8 to 20 positive axillary nodes.
- Age group of 39yrs to 65yrs have shown more number of patients with 20+ positive axillary nodes, most of which have resulted in death.

##**Pair Plots**

- 3C2 = 3 unique graphs will be obtained since Haberman dataset has 3 features.

[ ]:
```
mtplot.close()
sns.set_style('whitegrid')
```

```
sns.pairplot(haberman, hue='status', height=4).fig.suptitle('Pairplot Analysis␣
 ↪of Haberman Dataset', y=1.08)
mtplot.show()
```



Observations

No lines can be drawn to differentiate between two features meaning there are too many overlaps and thus "if-else" statement on the basis of shady revealations may give inaccurate output.

Breast cancer can be seen among women of 30 or 30+ yrs. age.

1. Plot age vs nodes Analysis,

1. Patients of age group 30 to 40yrs show least number of death cases and number of detected

axillary nodes range between 0 to 15, with most being in the range 0f 0 to 8.

2. Patients having age between 39 to 50yrs. have frequent death cases although their total number of nodes range from 1 to 10.
3. Patients having age between 50(approx.) to 60yrs and number of detected nodes in the range of 0 to 50 with most patients having nodes number between 0 to 10, have good number survived cases.

4. Patients of age group 60 t0 80yrs show mixed number of death and survival cases and the total number of nodes detected in them range between 1 to 10.Age group of 60 to 66 among them show good number of death cases although positive axillary nodes detected in them range between 1 to 5.

Conclusion for age vs nodes plot

1. Breast cancer may occur in girls of age group 30 to 80yrs.
2. Most number girls with breast cancer have axillary nodes ranging from 1 to 5.Moderate number have positive axillary nodes ranging from 5 to 10.Few have positive axillary nodes ranging from 10 to 20.Fewer have 20+ positive axillary nodes
3. Although chance of survival increases with less number of nodes but people of age group 40 to 45 yrs and 60 to 66 yrs have shown high death cases inspite of having less number of positive axillary nodes rangine from 1 to 5 and people of age group 30 to 40 yrs have survived even with 1 to 30 positive axillary nodes.

2. Plot year vs age Analysis,

1. In year 1958 to 1960, more number of death cases were seen among age group 44 to 54 yrs and 60 to 70yrs.
2. In year 1961 to 1965, most death cases prevailed among age group of 43 to 68yrs.
3. In year 1965 to 1970, although death cases were seen among age group 35 to yrs., frequency of death is much lesser than than of 1958 to 1965.

Conclusion for year vs age plot

1. Death number was more in the initial years of the study and improved in the last 5 yrs of study i.e. between 1965 to 1970.
2. Age group of 41 yrs to 70 yrs were most affected.
3. 70+ yrs age group and 30 to 40yrs has shown lesser number death cases.

3. Plot year vs nodes Observation,

1. Death due to 1 to 25 positive axillary nodes is more in total duration of study.
2. Women having 1 to 10 positive axillary nodes have shown more survives cases between year 1960 to 1964.

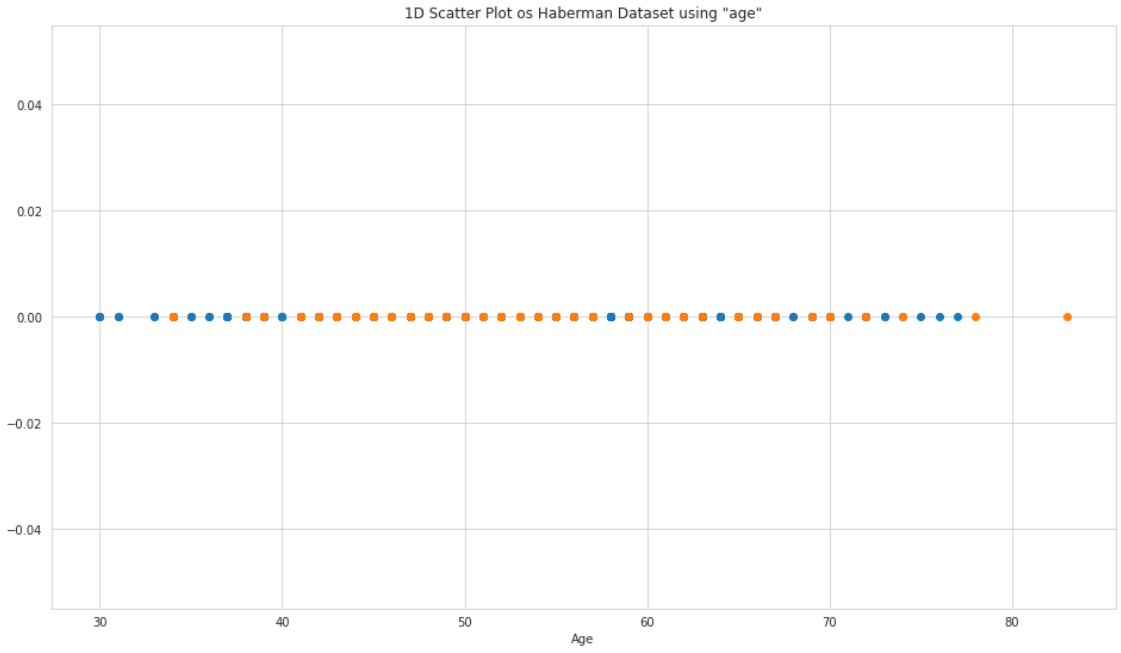
3.1.1 Conclusion: Perfect features combination can be (age vs nodes) because it gives more relevant information to find out survival status of patients.

4 Univariate Exploratory Data Analysis

4.1 1D Scatter Plot

4.1.1 Histogram

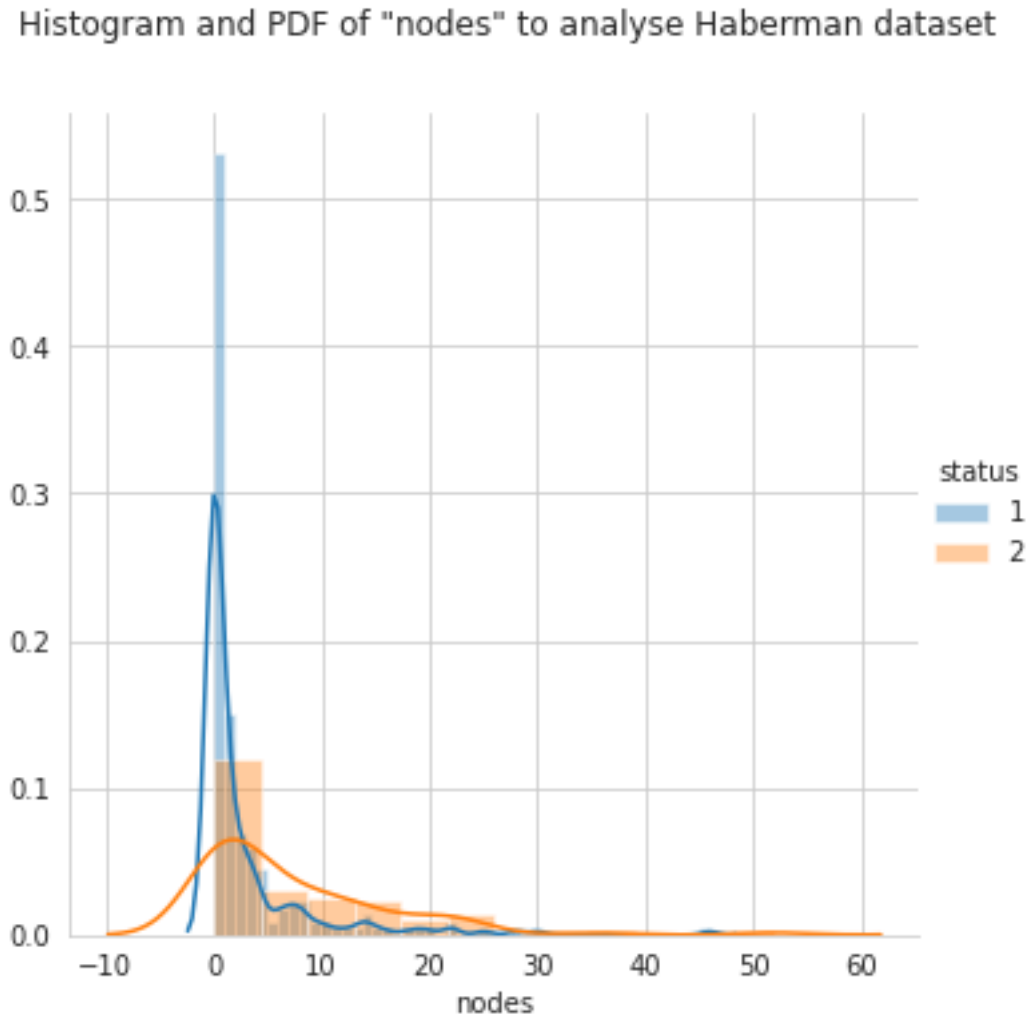
[ ]:
```
# 1D Scatter Plot of age

survived = haberman.loc[haberman['status'] == 1]
died = haberman.loc[haberman['status'] == 2]

mtplot.plot(survived['age'],np.zeros_like(survived['age']), 'o')
mtplot.plot(died['age'], np.zeros_like(died['age']), 'o')
mtplot.title('1D Scatter Plot os Haberman Dataset using "age"')
mtplot.xlabel('Age')
mtplot.show()
```



Observation

- Age of people in range 38(approx.) to 76(approx.) are having more death cases but again survival cases might have been overlapped by the dead cases.
- Age of people in range 30 to 36yrs have shown lesser death cases.
- Since there are too many overlaps of dead and survived patients data, conclusions drawn can be clearly inaccurate.

```
[ ]: # Histogram and PDF to analyse Haberman dataset
sns.FacetGrid(haberman, hue = 'status', height = 5).map(sns.distplot,'nodes').
 ↪add_legend().fig.suptitle('Histogram and PDF of "nodes" to analyse Haberman␣
 ↪dataset', y=1.08)
mtplot.show()
```

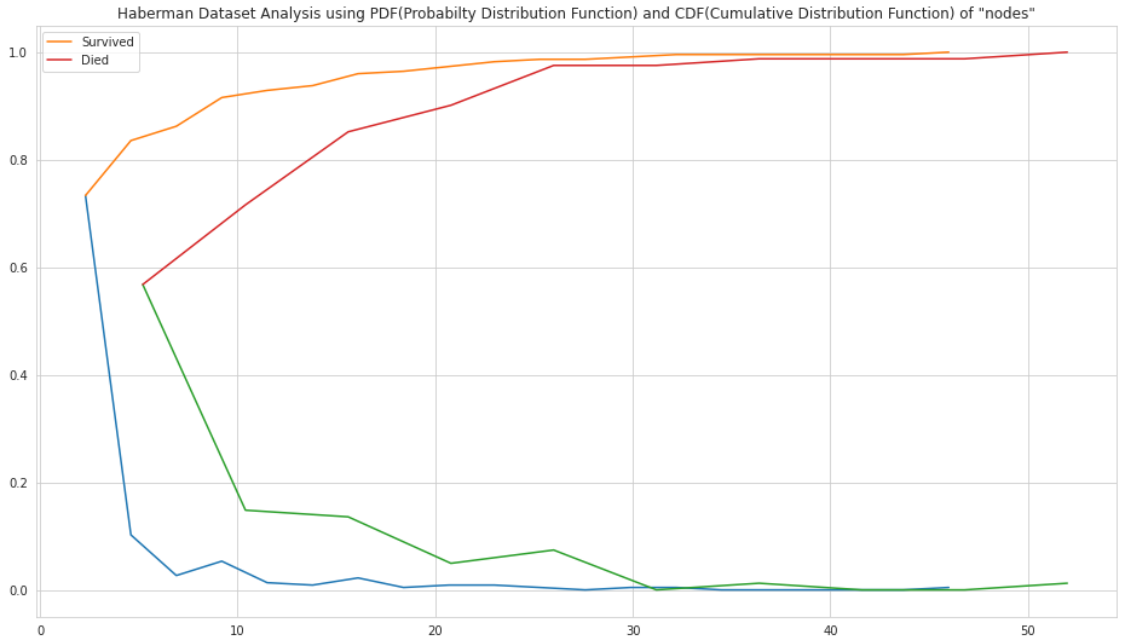


Observation:

1. Most death case are seen to have nodes between 0 to 5.
2. Most survived cases are seen to have nodes between 0 to 2.

4.1.2 Exploratory Data Analysis by Probabilty Density Function(PDF) and Cumulative Density Function(CDF)

[ ]:
```
#survived
counts, bin_edges = np.histogram(survived['nodes'], bins = 20, density = True)
pdf = counts/(sum(counts))
#print('pdf for survived["age"]: ',pdf)
#print('bin_edges for survived["age"]: ',bin_edges)
cdf = np.cumsum(pdf)
mtplot.plot(bin_edges[1:], pdf)
mtplot.plot(bin_edges[1:], cdf, label='Survived')
mtplot.legend(loc='center')

#died
counts, bin_edges = np.histogram(died['nodes'], bins = 10, density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
mtplot.plot(bin_edges[1:], pdf)
mtplot.plot(bin_edges[1:], cdf, label='Died')
mtplot.legend(loc='best')

mtplot.rcParams["figure.figsize"] = [16,9]
mtplot.title('Haberman Dataset Analysis using PDF(Probabilty Distribution␣
 ↪Function) and CDF(Cumulative Distribution Function) of "nodes"',␣
 ↪fontweight=10)
mtplot.show()
```



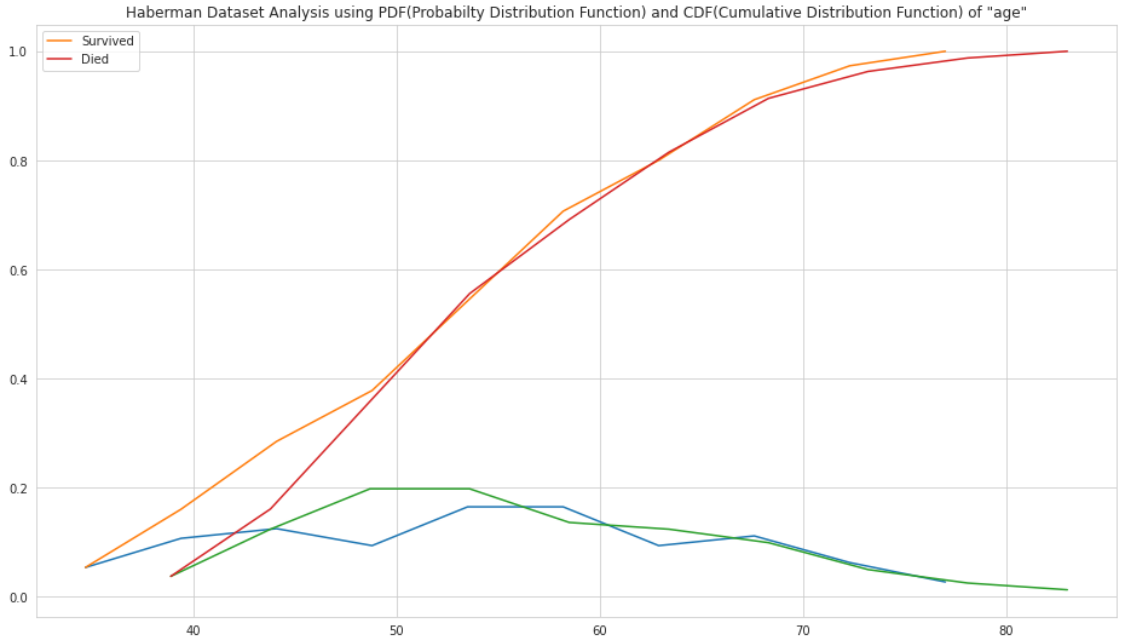
Observation

1. Women with (nodes <= 6), 85% of times patient will survive.
2. if (nodes > 6), 58% time patient will die

[ ]:
```
def status_is(nodes):
  if (nodes <= 6): # Correct result about 85% of time.
    print('Survived. Probablity is 85%.')
  elif (nodes > 6):
    print('Died. Probability is 58%.')

status_is(3)
```

```
Survived. Probablity is 85%.
```

[ ]:
```
#PDF, CDF analysis using age.

#survived
counts, bin_edges = np.histogram(survived['age'], bins = 10, density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
mtplot.plot(bin_edges[1:], pdf)
mtplot.plot(bin_edges[1:], cdf, label='Survived')
mtplot.legend()

#Died
counts, bin_edges = np.histogram(died['age'], bins=10, density=True)
pdf=counts/(sum(counts))
cdf=np.cumsum(pdf)
mtplot.plot(bin_edges[1:], pdf)
mtplot.plot(bin_edges[1:], cdf, label='Died')
mtplot.legend()

mtplot.title('Haberman Dataset Analysis using PDF(Probabilty Distribution␣
 ↪Function) and CDF(Cumulative Distribution Function) of "age"', fontweight=10)
mtplot.show()
```



Observation

1. Women with age $<$ 38yrs have 18% chances of survival.

4.1.3 Exploratory Data Ananlysis using Mean, Variance and Standard Deviation

[ ]:
```
#Mean of class label using Nodes
print('Mean of class labels using Nodes, ')
print('\tMean of Survived cases: ', (np.mean(survived['nodes'])))
print('\tMean of Death cases: ', (np.mean(died['nodes'])))

print(' ')
#Mean of class label using Age
print('Mean of class labels using Age, ')
print('\tMean of Survived cases: ', (np.mean(survived['age'])))
print('\tMean of Death cases: ', (np.mean(died['age'])))

print(' ')
#Mean of class label using Year
print('Mean of class labels using Year, ')
print('\tMean of Survived cases: ', (np.mean(survived['year'])))
print('\tMean of Death cases: ', (np.mean(died['year'])))
```

```
Mean of class labels using Nodes,
        Mean of Survived cases:  2.7911111111111113
        Mean of Death cases:  7.45679012345679
```

```
Mean of class labels using Age,
        Mean of Survived cases:  52.01777777777778
        Mean of Death cases:  53.67901234567901

Mean of class labels using Year,
        Mean of Survived cases:  62.86222222222222
        Mean of Death cases:  62.82716049382716
```

Observation

1. Patients who died within 5 years have much higher number of nodes than the patient who survived for 5 years or longer after operation.
2. Patients who have survived and died have almost same mean age.
3. Year of survived and death cases have same average year.

[ ]:
```
#Standard Deviation of Class Label using Nodes
print('Standard Deviation of Class Label using Nodes,')
print('\tStandard Deviation of Survived cases: ', np.std(survived['nodes']))
print('\tStandard Deviation of Death cases: ', np.std(died['nodes']))

print(' ')
#Standard Deviation of Class Label using Age
print('Standard Deviation of Class Label using Age,')
print('\tStandard Deviation of Survived cases: ', np.std(survived['age']))
print('\tStandard Deviation of Death cases: ', np.std(died['age']))

print(' ')
#Standard Deviation of Class Label using Year
print('Standard Deviation of Class Label using Year,')
print('\tStandard Deviation of Survived cases: ', np.std(survived['year']))
print('\tStandard Deviation of Death cases: ', np.std(died['year']))
```

```
Standard Deviation of Class Label using Nodes,
        Standard Deviation of Survived cases:  5.857258449412131
        Standard Deviation of Death cases:  9.128776076761632

Standard Deviation of Class Label using Age,
        Standard Deviation of Survived cases:  10.98765547510051
        Standard Deviation of Death cases:  10.10418219303131

Standard Deviation of Class Label using Year,
        Standard Deviation of Survived cases:  3.2157452144021956
        Standard Deviation of Death cases:  3.3214236255207883
```

Observation:

1. Standard deviation of death cases is much higher than that of survived cases. Thus, spread of of death cases is higher than the spread of survived cases with respect to nodes. One conclusion can be drawn, women with higher number of detected nodes have died.
2. Age and year have nearly same standard deviation for death and survived cases respectively.

4.1.4 Exploratory Data Ananlysis using Median, Percentile, Quantile, IQR, MAD

[ ]:
```
#Median of Class Label using Nodes
print('Median of Class Label using Nodes, ')
print('\tMedian of Survived Cases: ', np.median(survived['nodes']))
print('\tMedian of Death Cases: ', np.median(died['nodes']))

print(' ')
#Median of Class Label using Age
print('Median of Class Label using Age, ')
print('\tMedian of Survived Cases: ', np.median(survived['age']))
print('\tMedian of Death Cases: ', np.median(died['age']))

print(' ')
#Median of Class Label using Year
print('Median of Class Label using Year, ')
print('\tMedian of Survived Cases: ', np.median(survived['year']))
print('\tMedian of Death Cases: ', np.median(died['year']))
```

```
Median of Class Label using Nodes,
        Median of Survived Cases:  0.0
        Median of Death Cases:  4.0

Median of Class Label using Age,
        Median of Survived Cases:  52.0
        Median of Death Cases:  53.0

Median of Class Label using Year,
        Median of Survived Cases:  63.0
        Median of Death Cases:  63.0
```

Observation:

1. Median of death cases is much higher than that of survived cases.
2. Age and year have nearly same and same median for death and survived cases respectively.

[ ]:
```
#Quantile of Class Label using Nodes
print('Quantile of Class Label using Nodes, ')
print('\tQuantiles of Survived cases: ',(np.percentile(survived['nodes'], np.
 ↪arange(0, 100, 25))))
print('\tQuantiles of death cases: ',(np.percentile(died['nodes'], np.arange(0,␣
 ↪100, 25))))

print(' ')
#Quantile of Class Label using Age
print('Quantile of Class Label using Age, ')
print('\tQuantiles of Survived cases: ',(np.percentile(survived['age'], np.
 ↪arange(0, 100, 25))))
```

```
print('\tQuantiles of death cases: ',(np.percentile(died['age'], np.arange(0,␣
 ↪100, 25))))

print(' ')
#Quantile of Class Label using Year
print('Quantile of Class Label using Year, ')
print('\tQuantiles of Survived cases: ',(np.percentile(survived['year'], np.
 ↪arange(0, 100, 25))))
print('\tQuantiles of death cases: ',(np.percentile(died['year'], np.arange(0,␣
 ↪100, 25))))
```

```
Quantile of Class Label using Nodes,
        Quantiles of Survived cases:  [0. 0. 0. 3.]
        Quantiles of death cases:  [ 0.  1.  4. 11.]

Quantile of Class Label using Age,
        Quantiles of Survived cases:  [30. 43. 52. 60.]
        Quantiles of death cases:  [34. 46. 53. 61.]

Quantile of Class Label using Year,
        Quantiles of Survived cases:  [58. 60. 63. 66.]
        Quantiles of death cases:  [58. 59. 63. 65.]
```

[ ]:
```
# 90th Percentile for Class Label of Nodes
print('90th Percentile for Class Label of Nodes, ')
print('\t90th Percentile of Survived cases: ', np.percentile(survived['nodes'],␣
 ↪90))
print('\t90th Percentile of Death Cases: ', np.percentile(survived['nodes'],␣
 ↪90))

# 90th Percentile for Class Label of Age
print('\n90th Percentile for Class Label of Age, ')
print('\t90th Percentile of Survived cases: ', np.percentile(survived['age'],␣
 ↪90))
print('\t90th Percentile of Death Cases: ', np.percentile(died['age'], 90))

# 90th Percentile for Class Label of Year
print('\n90th Percentile for Class Label of Year, ')
print('\t90th Percentile of Survived cases: ', np.percentile(survived['year'],␣
 ↪90))
print('\t90th Percentile of Death cases: ', np.percentile(died['year'], 90))
```

```
90th Percentile for Class Label of Nodes,
        90th Percentile of Survived cases:  8.0
        90th Percentile of Death Cases:  8.0

90th Percentile for Class Label of Age,
```

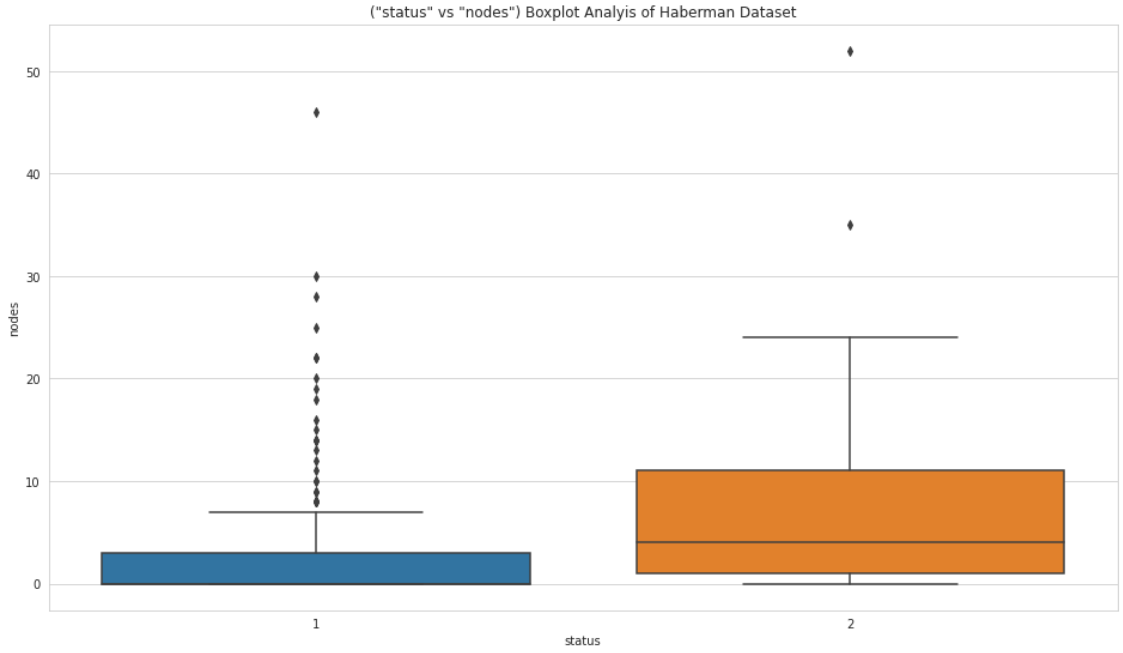
```
        90th Percentile of Survived cases:  67.0
        90th Percentile of Death Cases:  67.0

90th Percentile for Class Label of Year,
        90th Percentile of Survived cases:  67.0
        90th Percentile of Death cases:  67.0
```

[ ]:
```
from statsmodels import robust
# Mean Absolute Deviation for class Label of Nodes
print('Mean Absolute Deviation of Nodes')
print('\tMean Absolute Deviation survived cases: ', robust.
 ↪mad(survived['nodes']))
print('\tMean Absolute Deviation death cases: ', robust.mad(died['nodes']))

# Mean Absolute Deviation for class Label of Age
print('Mean Absolute Deviation of Age')
print('\tMean Absolute Deviation survived cases: ', robust.mad(survived['age']))
print('\tMean Absolute Deviation death cases: ', robust.mad(died['age']))

# Mean Absolute Deviation for class Label of Year
print('Mean Absolute Deviation of Year')
print('\tMean Absolute Deviation survived cases: ', robust.
 ↪mad(survived['year']))
print('\tMean Absolute Deviation death cases: ', robust.mad(died['year']))
```

```
Mean Absolute Deviation of Nodes
        Mean Absolute Deviation survived cases:  0.0
        Mean Absolute Deviation death cases:  5.930408874022408
Mean Absolute Deviation of Age
        Mean Absolute Deviation survived cases:  13.343419966550417
        Mean Absolute Deviation death cases:  11.860817748044816
Mean Absolute Deviation of Year
        Mean Absolute Deviation survived cases:  4.447806655516806
        Mean Absolute Deviation death cases:  4.447806655516806
```

###Box Plot

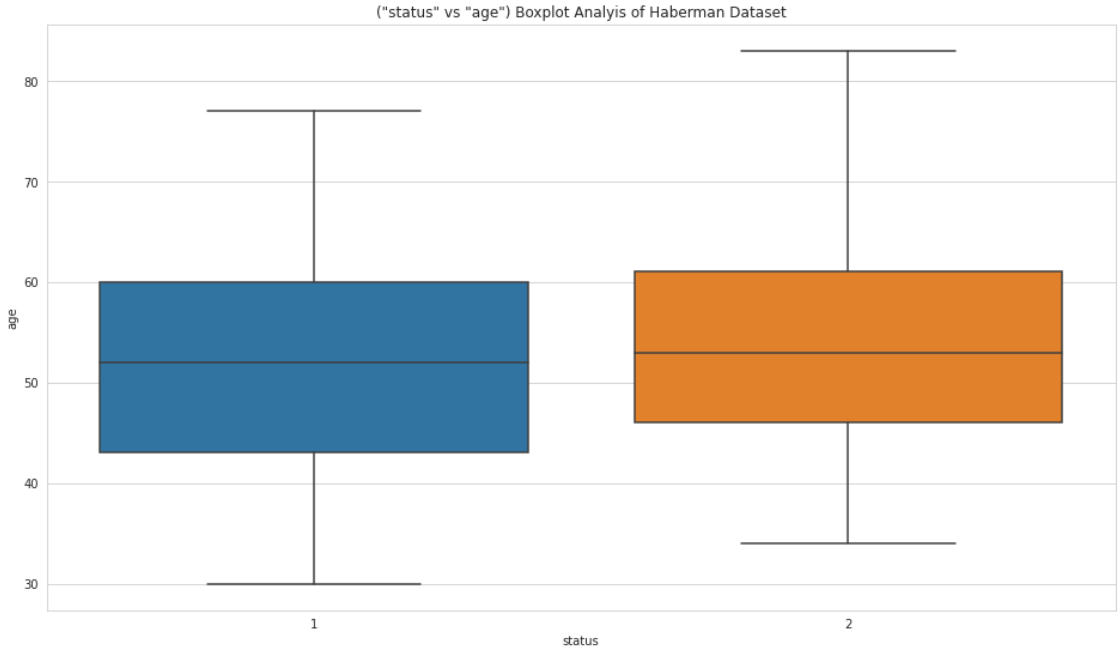
[ ]:
```
sns.boxplot(x = 'status', y = 'nodes', data = haberman)
mtplot.title('("status" vs "nodes") Boxplot Analyis of Haberman Dataset')
mtplot.show()
```



Observation

1. About 50% of women who have died had 1 to 11 positive axillary nodes.
2. Out of which 25% women who died haad 3 to 11 positive axillary nodes and rest 25% had 0 to 2 positive Axillary nodes.
3. 25% women who survived for 5 years or longer after operation had been detected with 0 to 2 to positive axillary nodes.
4. Women who had nodes > 17 had resulted in dying within 5years of operation.

[ ]:
```
sns.boxplot(x = 'status', y = 'age', data = haberman)
mtplot.title('("status" vs "age") Boxplot Analyis of Haberman Dataset')
mtplot.show()
```

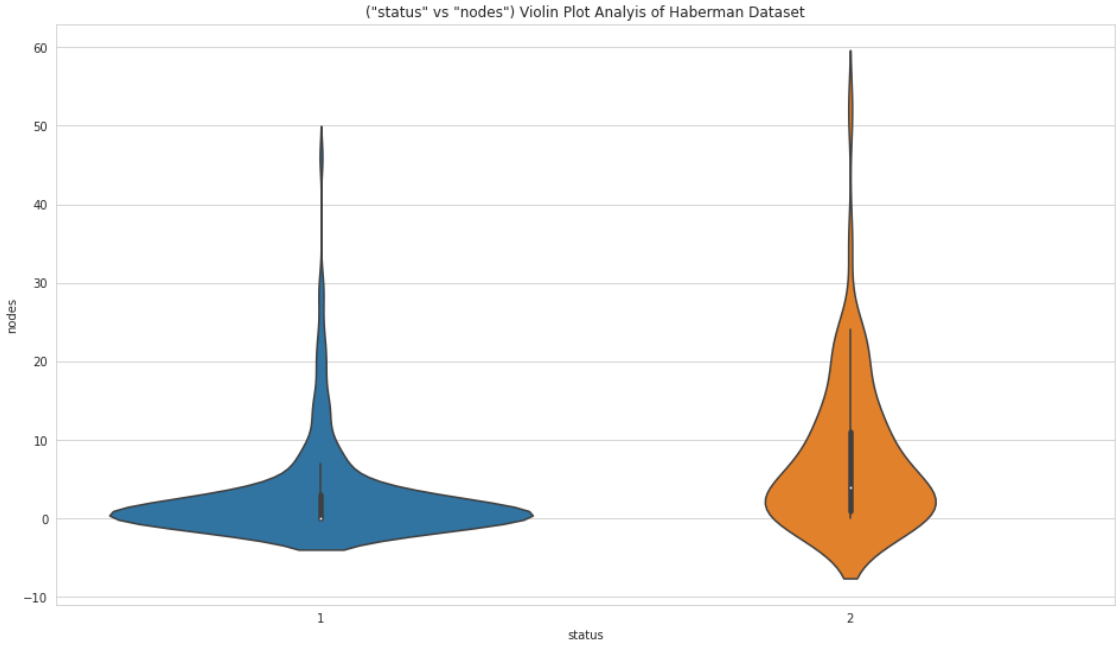


Observation

1. Women with age < 34 yrs have survived 5 years or longer after operation.
2. Women with age > 78 yrs died within 5 years after operation.

###Violin Plot

[ ]:
```
sns.violinplot(x = 'status', y = 'nodes', data = haberman, height = 10)
mtplot.title('("status" vs "nodes") Violin Plot Analyis of Haberman Dataset')
mtplot.show()
```



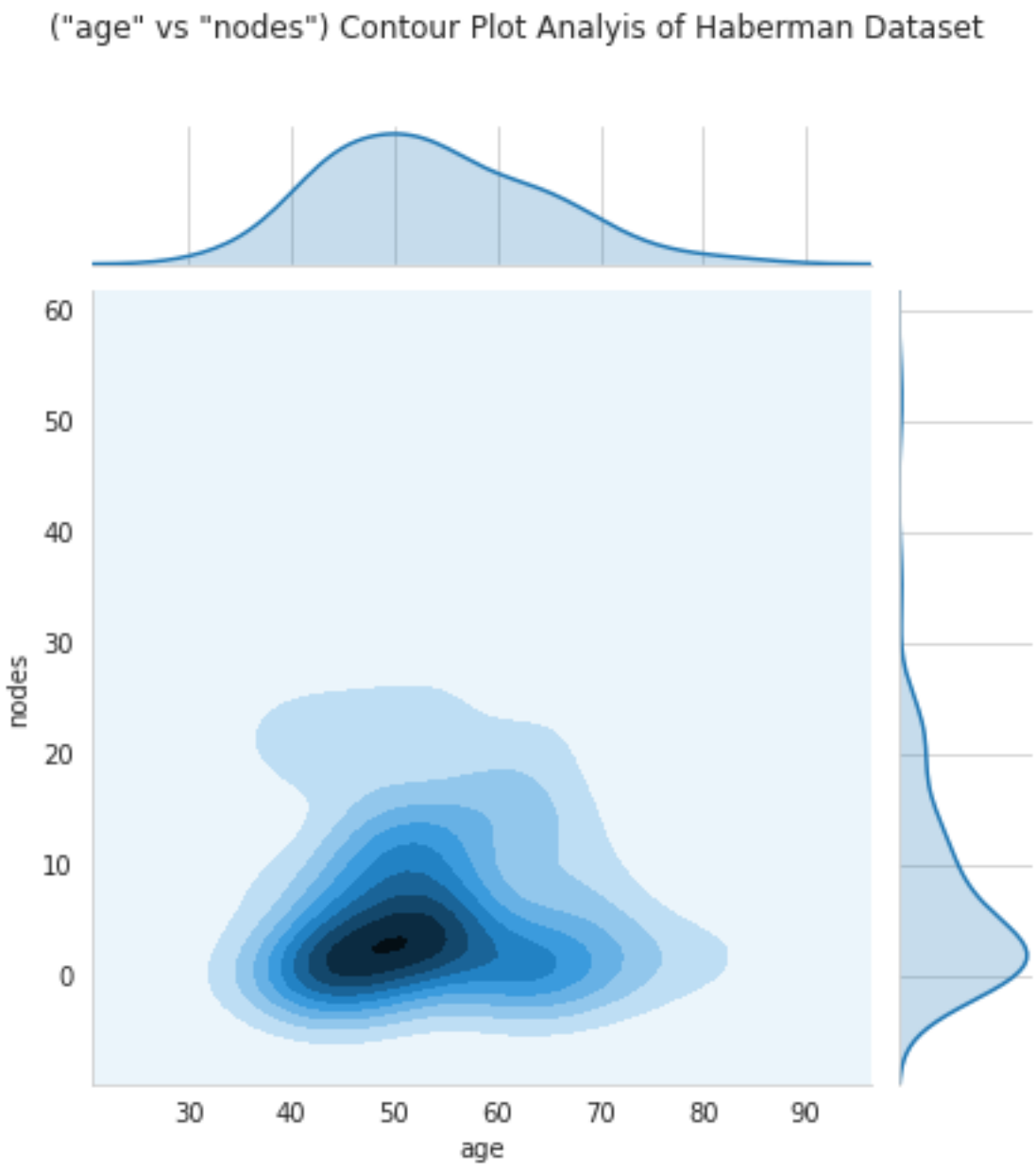
Observation

1. 50% women who survived for 5 yrs or longer had 0 to 1 postive axillary nodes after Operation.
2. 75% women who survived for 5yrs or longer after operation had 0 to 3 positive axillary nodes.
3. 50% women who died within 5 yrs after operation had 0 to 4 positive axillary nodes.
4. 75% women who died within 5 yrs after operation had 0 to 12 positive axillary nodes.

5 Multivariate Analysis

5.1 Conrour Plot

[ ]:
```
sns.jointplot(x = 'age', y = 'nodes', data = died, kind = 'kde').fig.
 ↪suptitle('("age" vs "nodes") Contour Plot Analyis of Haberman Dataset', y=1.
 ↪08)
#mtplot.title('("status" vs "nodes") Boxplot Analyis of Haberman Dataset')
mtplot.show()
```



Observation

1. Women of age 48yrs to 51yrs have shown highest death cases with 1 to 3 detected positive axillary nodes.
2. Women of age group 43 years to 55 years have shown second highest death cases with almost 0 to 6 nodes.

6 Conclusion

1. Patients who survived more than 5 years or longer are more in number than the patients who died within 5 years of surgery.
2. Breast cancer can happen to women of age above 30yrs.
3. No line of separation could be drawn between to class label to differentiate between the two. Datapoints of both the class label had much overlapping to draw a line of separation.
4. Most women with breast cancer have detected postive axillary nodes ranging from 1 to 5.Moderate number have positive axillary nodes ranging from 5 to 10.Few have positive axillary nodes ranging from 10 to 20.Fewer have 20+ positive axillary nodes
5. Death cases were more in the initial years of the study and improved in the last 5 yrs of study i.e. between 1965 to 1970.
6. Women detected with less tha 6 nodes have 85% of survival and women detected with more than 6 nodes have 42% of survival.
7. Women with age $<$ 38yrs are seen to survive most.Higher risk of death are seen to be in women of age 40 to 60 years.
8. Above 60 yrs women who died have lower density with nodes between 0 to 5.