manojkrishnamoorthy87@gmail.com_1

August 2, 2019

1 EDA on Haberman's Dataset

Number of Data points:(306, 4) Independant variables:age, year, nodes Dependant variables or Labels:status

2 Haberman's Dataset

Data Description: The Habermans survival data set contains cases from a study that was conducted between 1958 and 1970 at the University of Chicagos Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Attribute Information: Age of patient at time of operation (numerical) Patients year of operation (year??1900, numerical) Number of positive auxillary nodes detected (numerical) Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the patient died within 5 years

OBJECTIVE: To perform EDA on data to investigate the given data through graphical representation. Classify survival status of a patient based on age, year of operation and number of nodes detected.

```
In [8]: import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        import numpy as np
        haberman = pd.read_csv("haberman.csv")
```

```
In [9]: #Data Points
        print(haberman.shape)
```

```
(306, 4)
```

```
In [10]: #column names
         print(haberman.columns)
```

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

```
In [11]: haberman["status"].value_counts()
```

```
Out[11]: 1    225
         2     81
         Name: status, dtype: int64
```

The above values show that 225 people have survived for more than 5 years after treatment while 81 people have not. Also this is an imbalance data set, as the data points for each class varies widely.

```
In [12]: haberman.describe()
```

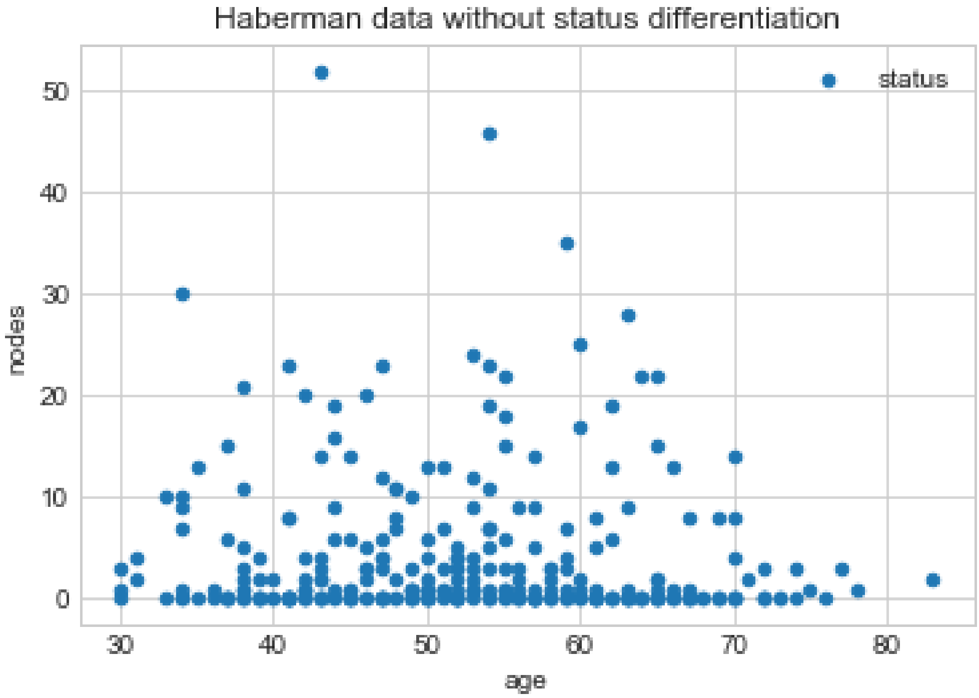
```
Out[12]:               age         year       nodes      status
         count  306.000000  306.000000  306.000000  306.000000
         mean    52.457516   62.852941    4.026144    1.264706
         std     10.803452    3.249405    7.189654    0.441899
         min     30.000000   58.000000    0.000000    1.000000
         25%     44.000000   60.000000    0.000000    1.000000
         50%     52.000000   63.000000    1.000000    1.000000
         75%     60.750000   65.750000    4.000000    2.000000
         max     83.000000   69.000000   52.000000    2.000000
```

```
In [13]: haberman['nodes'].max()
```

```
Out[13]: 52
```

3 2D-Scatter Plot

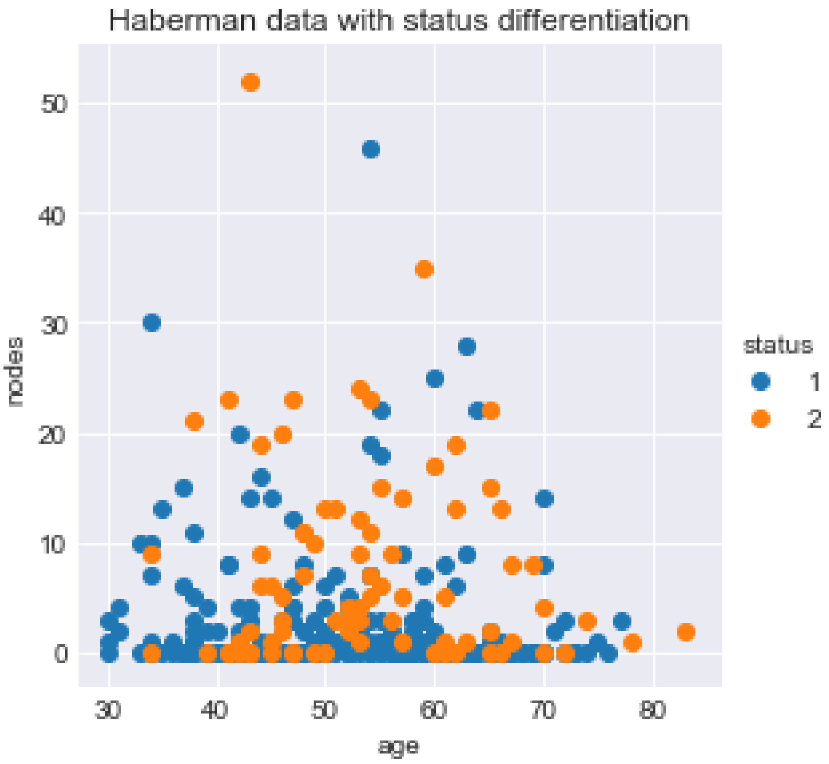
```
In [14]: #2D-scatter plot without using color for status differentiation
         haberman.plot(kind='scatter', x='age', y='nodes',label='status')
         plt.title('Haberman data without status differentiation')
         #plt.legend()
         plt.show()
```



Observation:

The above 2-D scatter plot is plotted with age and nodes on x and y axis respectively. The blue dots on the plot indicate the status of the person. Since both possible statuses are indicated by blue color alone, we can't infer much from the above plot.

```
In [15]: #2D-scatter plot using color for status differentiation
         sns.set_style("darkgrid")
         sns.FacetGrid(haberman, hue="status", size=4).map(plt.scatter, 'age', 'nodes').add_leg
         plt.title('Haberman data with status differentiation')
         plt.show();
```

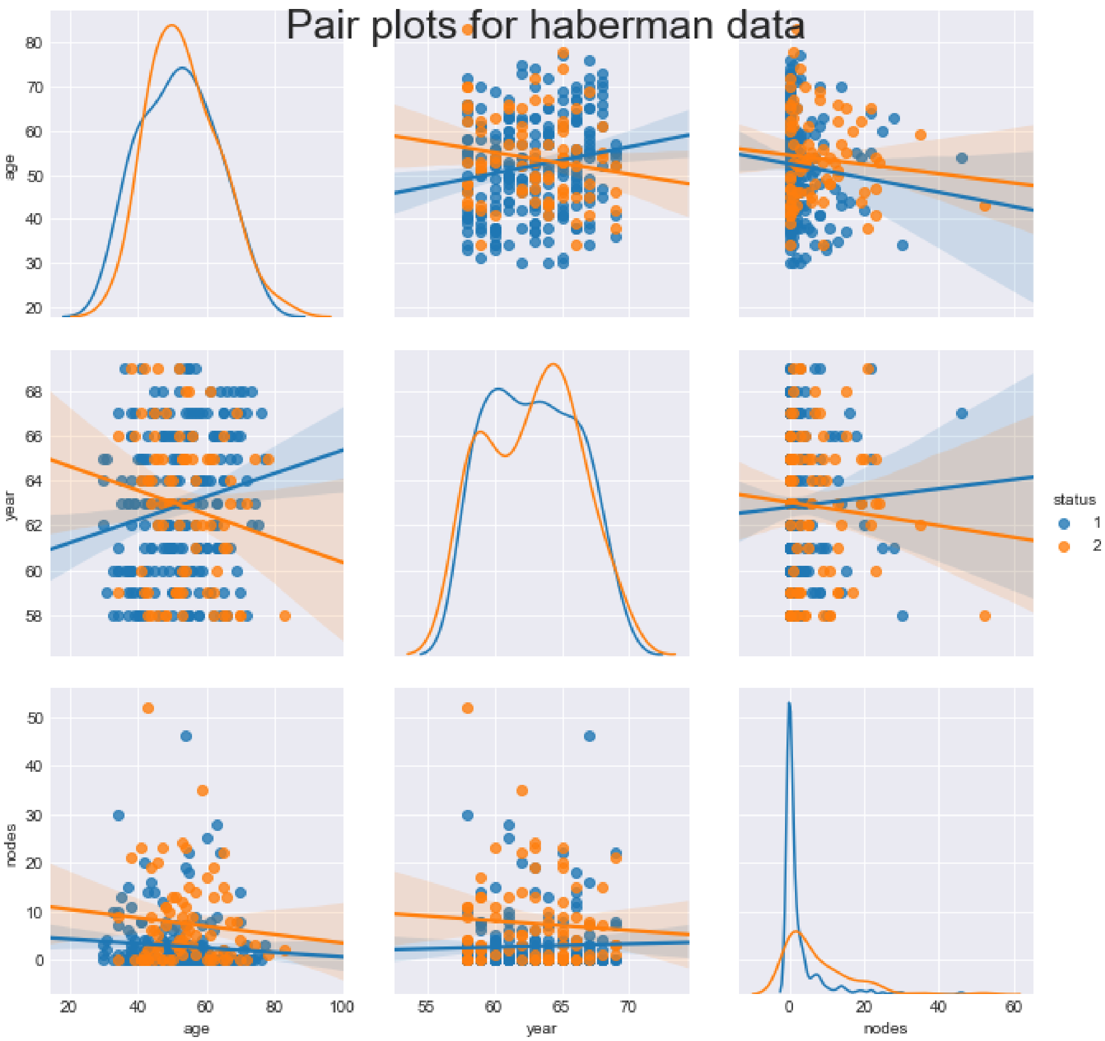


OBSERVATION:

Based on the above plot, it is impossible to linearly separate the patient's status using age and number of nodes. The color indicating both status are inseparable at any scale of age and number of nodes.

4 Pair Plot

```
In [16]: #haberman['status'] = haberman['status'].apply(lambda x: 'Positive' if x == 1 else 'N
         #print(haberman)
         #or
         plt.close();
         sns.pairplot(haberman,hue='status',vars=['age','year','nodes'], size=3, kind='reg', di
         plt.suptitle("Pair plots for haberman data", size=25)
         plt.show()
```

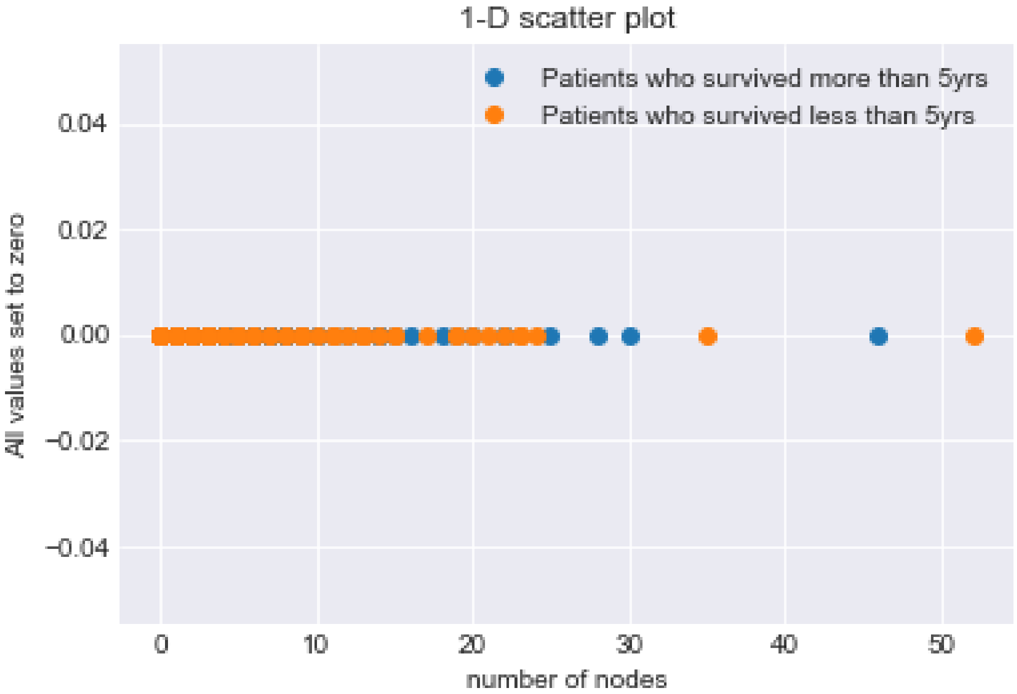


OBSERVATIONS: As the data points of both classes are overlapping each other, it is impossible to come to a conclusion on how to classify the patients status based on pair-plots above. But one observation that can be seen from the PDF of nodes is that patients with less number of nodes has better chance of surving more than 5 years. Also from the pair plot of age and nodes, we can infer young and middle-aged patients with less nodes have better chance of surviving for more than 5 years.

5 Histogram, PDF, CDF

6 1-D Scatter plots

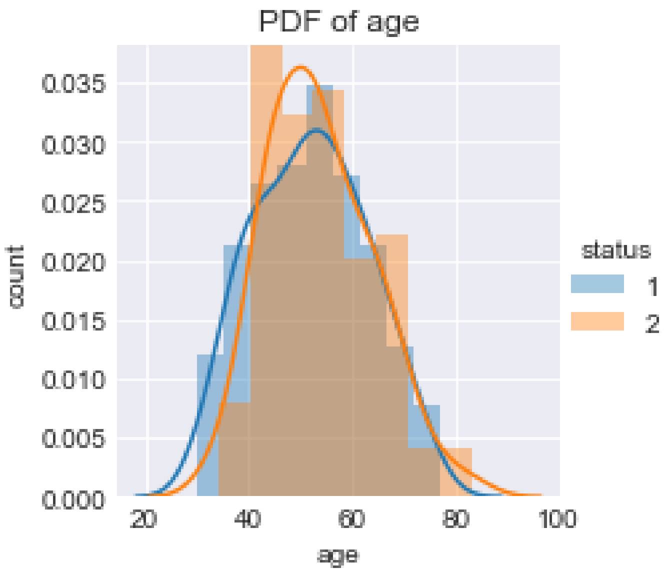
```
In [17]: haberman_st1=haberman.loc[haberman['status']==1]
         haberman_st2=haberman.loc[haberman['status']==2]
         plt.plot(haberman_st1["nodes"], np.zeros_like(haberman_st1['nodes']), 'o', label='Pat
         plt.plot(haberman_st2["nodes"], np.zeros_like(haberman_st2['nodes']), 'o', label='Pat
         plt.legend()
         plt.xlabel('number of nodes')
         plt.ylabel('All values set to zero')
         plt.title('1-D scatter plot')
         plt.show()
```



Observation: The above plot shows the density of the survival status of a patient with respect to the number of nodes. Since the scale is set to zero for all values on y-axis, the points overlap and we are unable to determine the density of points at a single place in plot.

7 Probablity Density Function

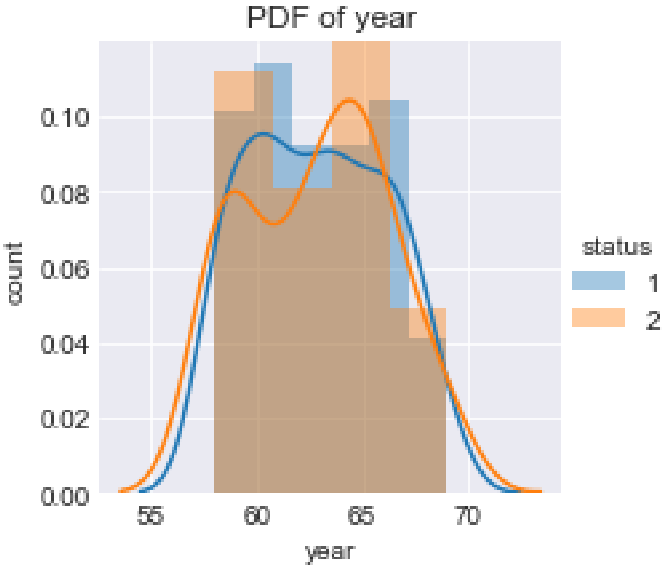
```
In [28]: import warnings
         warnings.filterwarnings("ignore")
         plt.close();
         sns.FacetGrid(haberman, hue='status', size=3).map(sns.distplot, 'age').add_legend()
         plt.title('PDF of age')
         plt.ylabel('count')
         plt.show()
```



Observation: From the above PDF obtained using age, we can clearly see both statuses of PDF overlapping each other. So we can't classify the survival status based on age.

```
In [29]: import warnings
         warnings.filterwarnings("ignore")
         sns.FacetGrid(haberman, hue='status', size=3).map(sns.distplot, 'year').add_legend()
         plt.title('PDF of year')
         plt.ylabel('count')
         plt.show
```

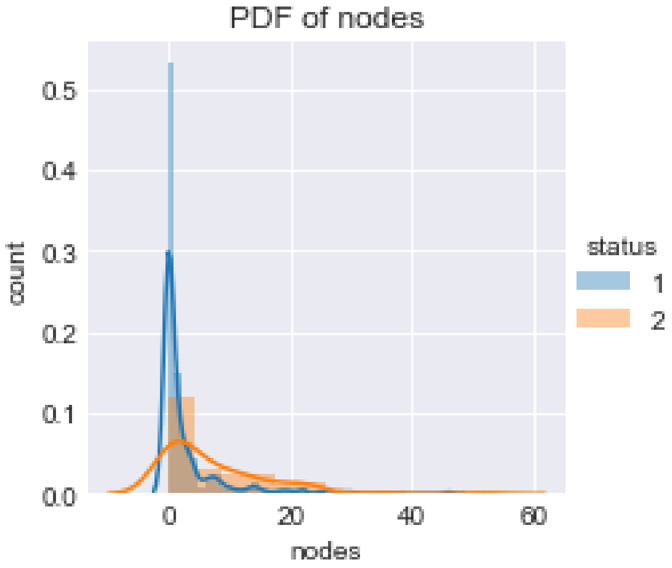
```
Out[29]: <function matplotlib.pyplot.show(*args, **kw)>
```



Observation: This more or less gives the same kind of inference obtained from the PDF of age plot. So, classification is impossible.

```
In [30]: import warnings
         warnings.filterwarnings("ignore")
         sns.FacetGrid(haberman, hue='status', size=3).map(sns.distplot, 'nodes').add_legend()
         plt.title('PDF of nodes')
         plt.ylabel('count')
         plt.show
```

```
Out[30]: <function matplotlib.pyplot.show(*args, **kw)>
```



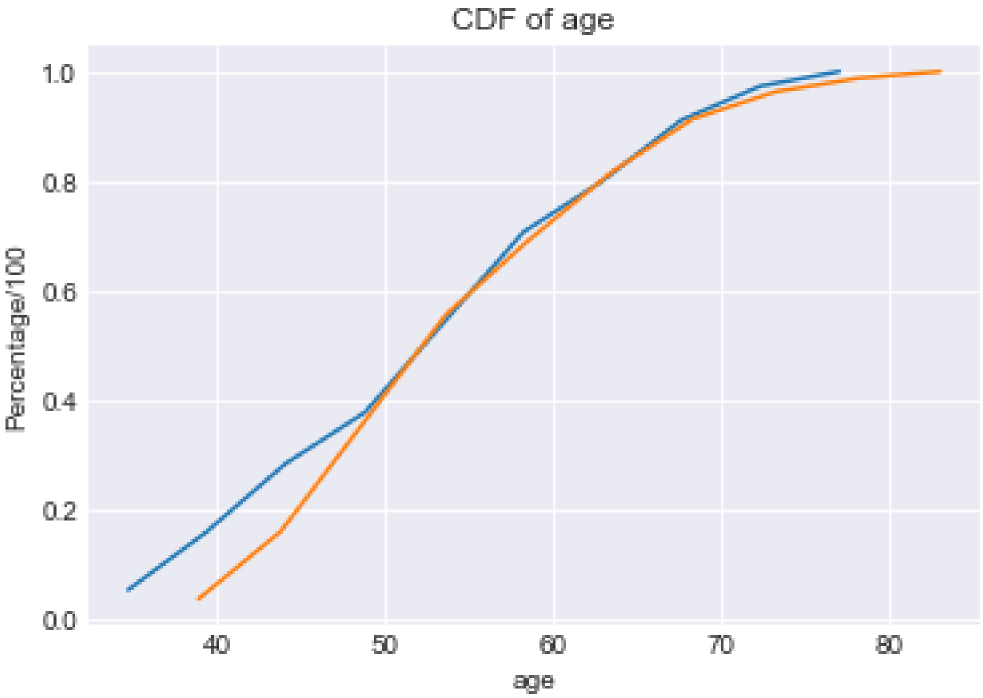
Observation: From the PDF of nodes we can infer that, the patients with less number of nodes has better chance of surving more than 5 years. As the number of nodes increases the comparitive rate of surviving more than 5 years decreases.

8 Cumulative Distribution Function

```
In [31]: counts, bin_edges = np.histogram(haberman_st1['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)
         plt.show

         counts, bin_edges = np.histogram(haberman_st2['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)
         plt.title('CDF of age')
         plt.xlabel('age')
         plt.ylabel('Percentage/100')
         plt.show
```

```
[0.05333333 0.10666667 0.12444444 0.09333333 0.16444444 0.16444444
 0.09333333 0.11111111 0.06222222 0.02666667]
[30.  34.7 39.4 44.1 48.8 53.5 58.2 62.9 67.6 72.3 77. ]
[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. ]
```

```
Out[31]: <function matplotlib.pyplot.show(*args, **kw)>
```

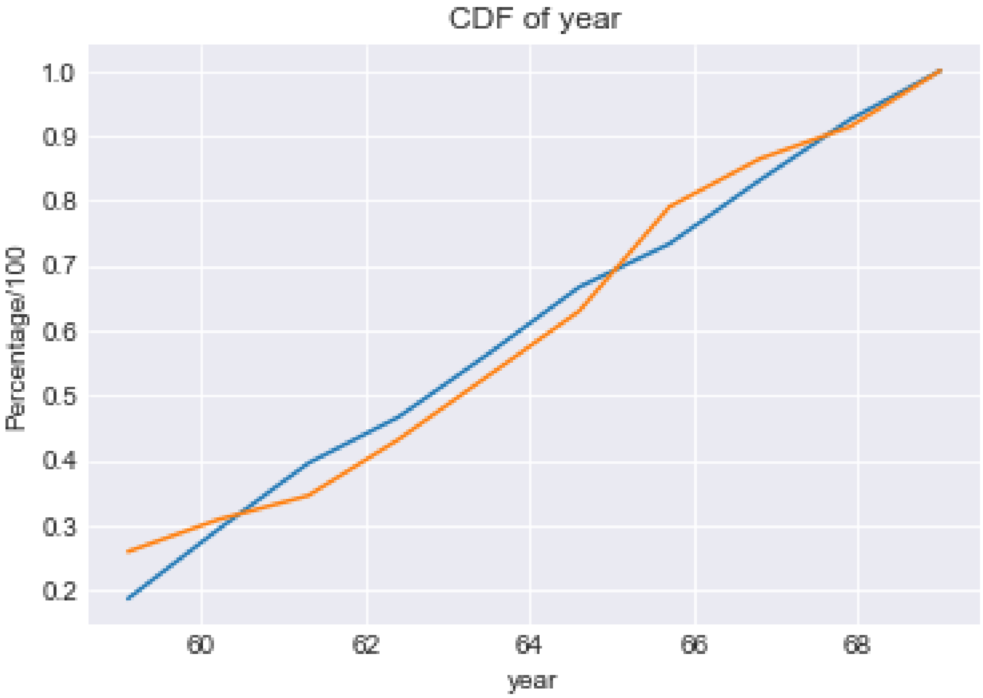


Observation: People over 76 years(approximately) has no possibility of surviving more than 5 years after treatment. The chance of surviving more than 5 years is more for people less than 50 years(approximately). Patients within 50 to 76 yrs have almost equal chance of surviving more than 5yrs.

In [32]:
```
counts, bin_edges = np.histogram(haberman_st1['year'], 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)
plt.show

counts, bin_edges = np.histogram(haberman_st2['year'], 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)
plt.title('CDF of year')
plt.xlabel('year')
plt.ylabel('Percentage/100')
plt.show
```

```
[0.18666667 0.10666667 0.10222222 0.07111111 0.09777778 0.10222222
 0.06666667 0.09777778 0.09333333 0.07555556]
[58.  59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
[0.25925926 0.04938272 0.03703704 0.08641975 0.09876543 0.09876543
 0.16049383 0.07407407 0.04938272 0.08641975]
[58.  59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
```

Out[32]:
```
<function matplotlib.pyplot.show(*args, **kw)>
```



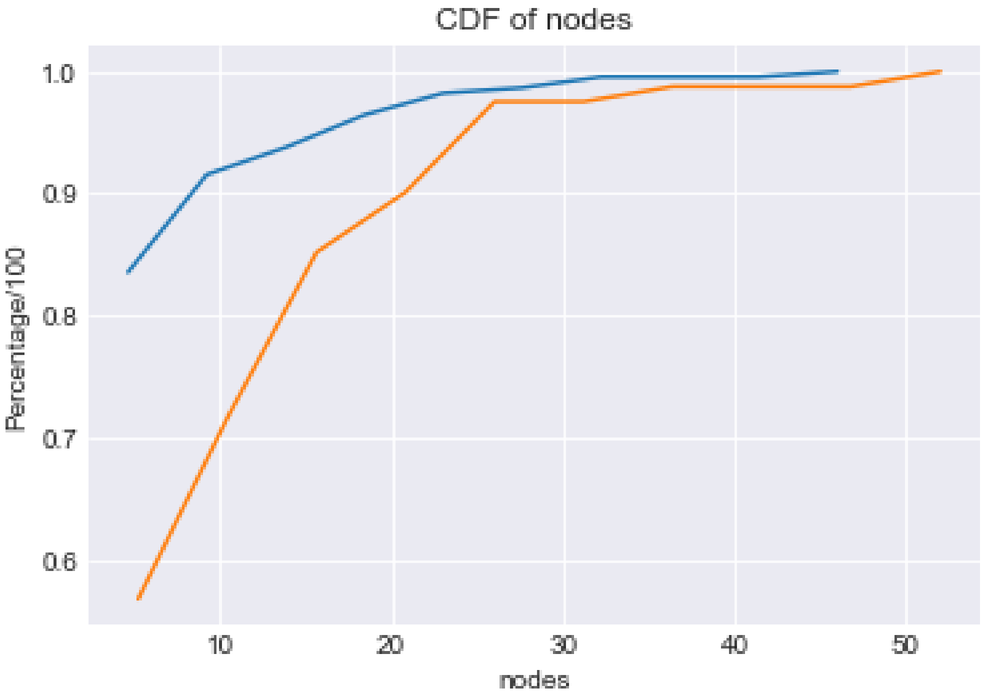
Observation: We can't infer much about the survival status from the year of operation.

```
In [33]: counts, bin_edges = np.histogram(haberman_st1['nodes'], 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)
         plt.show

         counts, bin_edges = np.histogram(haberman_st2['nodes'], 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)
         plt.title('CDF of nodes')
         plt.xlabel('nodes')
         plt.ylabel('Percentage/100')
         plt.show
```

```
[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. ]
[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. ]
```

```
Out[33]: <function matplotlib.pyplot.show(*args, **kw)>
```



OBSERVATION: People with less than 25 nodes(approx) has better survival rates(more than 5yrs). 98Patients with more than 46 nodes(approx) has almost no chance of surviving more than 5yrs after surgery.

9 Mean, Variance and Std-dev

```
In [53]: print("Means:")
         print("Average number of nodes of People who survived for more than 5yrs:")
         print(np.mean(haberman_st1["nodes"]))
         #Mean with an outlier.
         #print(np.mean(np.append(haberman_st1["nodes"],5000)));
         print("Average number of nodes of People who survived less than 5yrs:")
         print(np.mean(haberman_st2["nodes"]))
         print("\nStd-dev:");
         print("Standard Deviation of nodes of People who survived for more than 5yrs:")
         print(np.std(haberman_st1["nodes"]))
         print("Standard Deviation of nodes of People who survived less than 5yrs:")
         print(np.std(haberman_st2["nodes"]))
```

```
Means:
Average number of nodes of People who survived for more than 5yrs:
2.7911111111111113
Average number of nodes of People who survived less than 5yrs:
7.45679012345679

Std-dev:
Standard Deviation of nodes of People who survived for more than 5yrs:
5.857258449412131
Standard Deviation of nodes of People who survived less than 5yrs:
9.128776076761632
```

10 Median, Percentile, Quantile, IQR, MAD

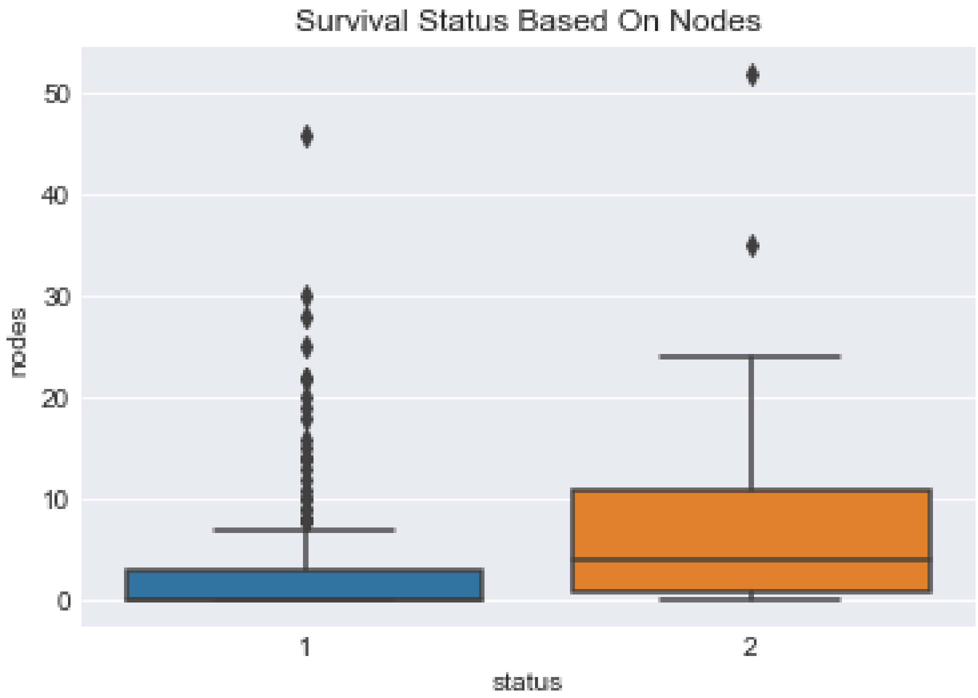
```
In [34]: print("\nMedians:")
         print("Average age of People who survived for more than 5yrs:")
         print(np.median(haberman_st1["age"]))
         #Median with an outlier
         #print(np.median(np.append(haberman_st1["age"],50000)));
         print("Average age of People who survived less than 5yrs:")
         print(np.median(haberman_st2["age"]))

         print("\nQuantiles:")
         print("Quartiles of  age of People who survived for more than 5yrs:")
         print(np.percentile(haberman_st1["age"],np.arange(0, 100, 25)))
         print("Quartiles of  age of People who survived less than 5yrs:")
         print(np.percentile(haberman_st2["age"],np.arange(0, 100, 25)))

         print("\n90th Percentiles:")
         print("90th Percentiles of  age of People who survived for more than 5yrs:")
         print(np.percentile(haberman_st1["age"],90))
         print("90th Percentiles of  age of People who survived less than 5yrs:")
         print(np.percentile(haberman_st2["age"],90))

         from statsmodels import robust
         print ("\nMedian Absolute Deviation")
         print("Median Absolute Deviation of  age of People who survived for more than 5yrs:")
         print(robust.mad(haberman_st1["age"]))
         print("Median Absolute Deviation of  age of People who survived less than 5yrs:")
         print(robust.mad(haberman_st2["age"]))
```

```
Medians:
Average age of People who survived for more than 5yrs:
52.0
Average age of People who survived less than 5yrs:
```

```
53.0

Quantiles:
Quartiles of  age of People who survived for more than 5yrs:
[30. 43. 52. 60.]
Quartiles of  age of People who survived less than 5yrs:
[34. 46. 53. 61.]

90th Percentiles:
90th Percentiles of  age of People who survived for more than 5yrs:
67.0
90th Percentiles of  age of People who survived less than 5yrs:
67.0

Median Absolute Deviation
Median Absolute Deviation of  age of People who survived for more than 5yrs:
13.343419966550417
Median Absolute Deviation of  age of People who survived less than 5yrs:
11.860817748044816
```

11 Box Plots

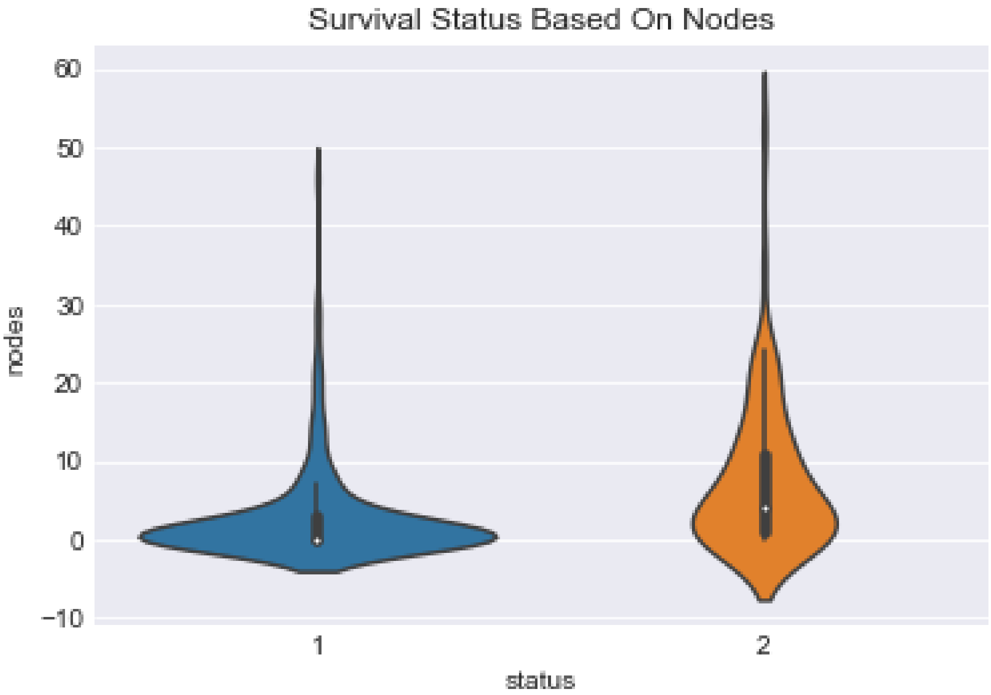
```
In [35]: sns.boxplot(x='status',y='nodes', data=haberman)
         plt.title('Survival Status Based On Nodes')
         plt.show()
```



OBSERVATION: Almost 75 percent of the patients who survived for more than 5yrs had less than 5 nodes. 50 percent of the patients who survived less than 5yrs had more than 5 nodes.

12 Violin Plot

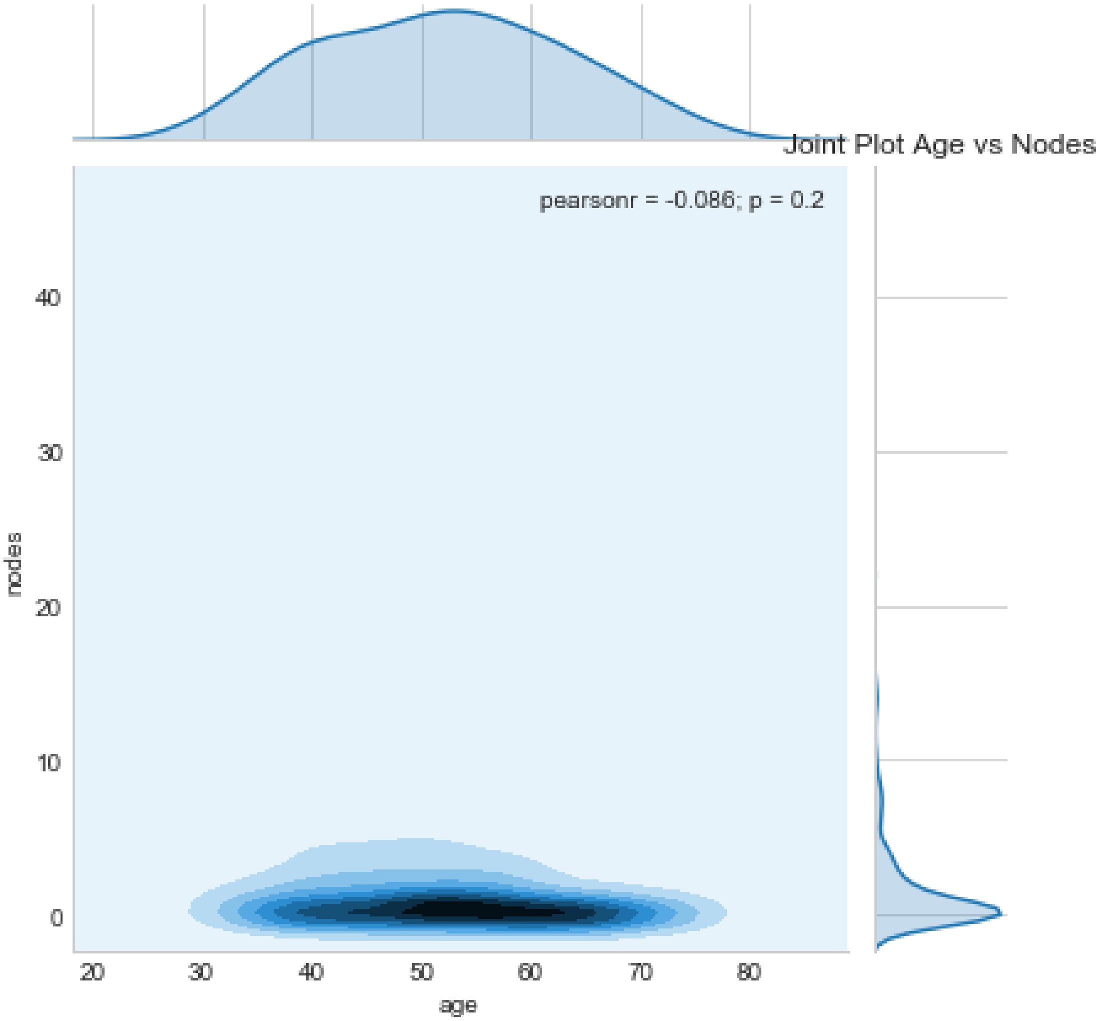
```
In [36]: sns.violinplot(x='status',y='nodes', data=haberman)
         plt.title('Survival Status Based On Nodes')
         plt.show()
```



Observation: The plot shows that the patients with less or no nodes have very high chance of surviving more than 5yrs. Survival rates(more than 5 years) decreases with increase in the node count.

13 Contour Plot

```
In [49]: import warnings
         warnings.filterwarnings("ignore")
         sns.jointplot(x="age", y="nodes", data=haberman_st1, kind="kde");
         plt.title('Joint Plot Age vs Nodes')
         plt.show();
```



Observation: The above shows the joint plot of patients with more than 5yrs survival. The count of patients with age 53 is more than the count of patients with any other age, who survived more than 5 years. Also people who had less than 4 nodes have a better chance of surviving more than 5 years.

14 Conclusion

1.Independant variables:age, year, nodes 2.Dependant variables or Labels:status 3.In the given data 225 people have survived for more than 5 years after treatment while 81 people have not.This is an imbalance data set, as the data points for each class varies widely. 4.From the PDF of nodes is that patients with less number of nodes has better chance of surving more than 5 years. Also from the pair plot of age and nodes, we can infer young and middle-aged patients with less nodes have better chance of surviving for more than 5 years. 5.People over 76 years(approximately) has no possibility of surviving more than 5 years after treatment. The chance of surviving more than 5 years is more for people less than 50 years(approximately). 6.Patients within 50 to 76 yrs have almost equal chance of surviving more than 5yrs. 7.People with less than 25 nodes(approx) has better survival rates(more than 5yrs). 988.Patients with more than 46 nodes(approx) has almost no chance of surviving more than 5yrs after surgery. 9.Almost 75 percent of the patients who survived for more than 5yrs had less than 5 nodes. 10.50 percent of the patients who survived less than 5yrs had more than 5 nodes. 11.The count of patients with age 53 is more than the count of patients with any other age, who survived more than 5 years. Also people who had less than 4 nodes have a better chance of surviving more than 5 years.