Habermans Dataset

March 25, 2018

1 Plotting for Exploratory Data Analysis(EDA) for Cancer Patients

2 Habermans Dataset

Sources: (a) Donor: Tjen-Sien Lim (b) Date: March 1999

Past Usage:

Haberman, S. J. (1976). Generalized Residuals for Log-Linear Models, Proceedings of the 9th International Biometrics Conference, Boston, pp. 104-122. Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984), Graphical Models for Assessing Logistic Regression Models (with discussion), Journal of the American Statistical Association 79: 61-83. Lo, W.-D. (1993). Logistic Regression Trees, PhD thesis, Department of Statistics, University of Wisconsin, Madison, WI. Relevant Information: 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.

- Number of Instances: 306
- Number of Attributes: 4 (including the class attribute)
- Attribute Information:
  - Age of patient at time of operation (numerical)
  - Patients year of operation (year - 1900, numerical)
  - Number of positive axillary nodes detected (numerical)
  - Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the patient died within 5 year
- Missing Attribute Values: None

3 Objective

Classify a new patient according to one of the 2 classes that is whether it survived 5 years or longer or patient died within 5 years, given the 3 features

```
In [1]: #importing all libraries
        import pandas as pd
        import seaborn  as se
        import numpy as np
        import matplotlib.pyplot as plt
```

```
In [79]: #reading the dataset
          hb = pd.read_csv("haberman.csv")
          #hb
```

```
In [3]: hb.shape
        #it shows we have 306 rows and 4 columns

Out[3]: (306, 4)
```

```
In [4]: hb.columns

Out[4]: Index(['Age', 'year', 'positive_axillary_nodes', 'survival_status'], dtype='object')
```

```
In [5]: hb['survival_status'].value_counts()

Out[5]: 1    225
        2     81
        Name: survival_status, dtype: int64
```

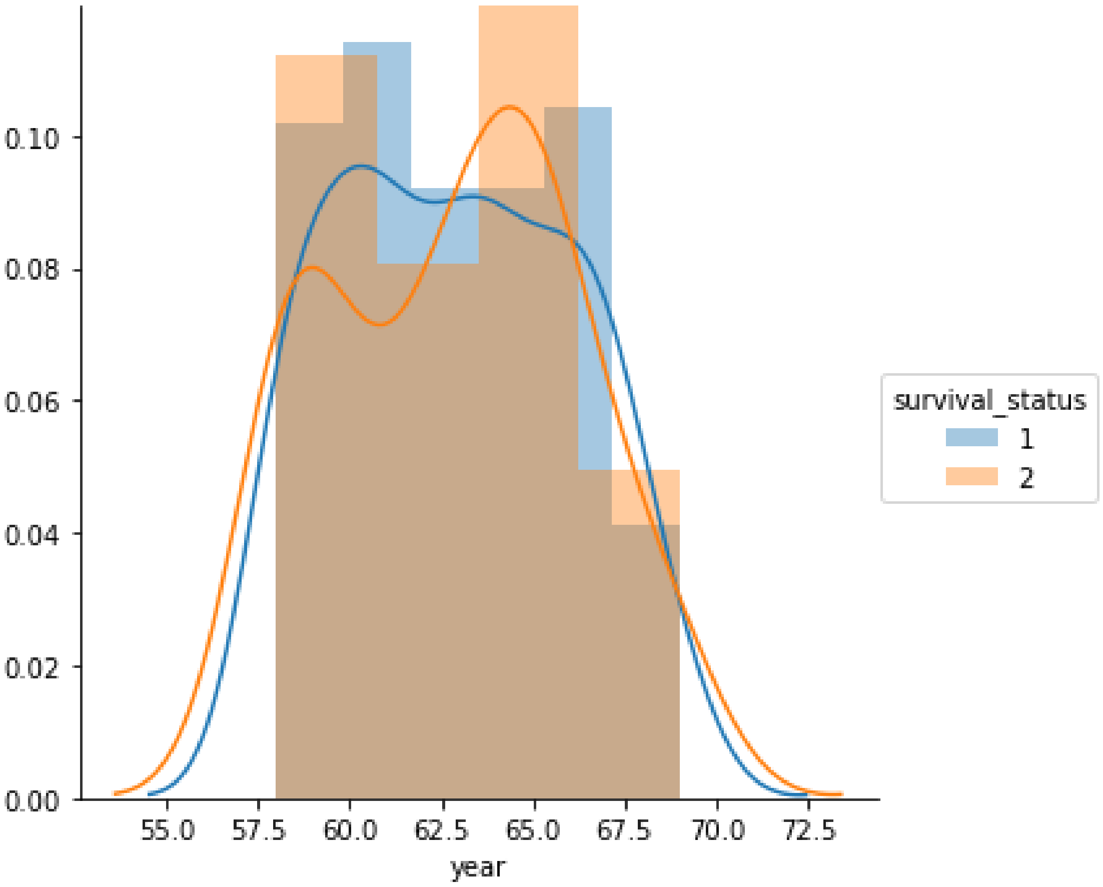
4 Observations

This shows * Only 225 patients survived 5 years or longer * And 81 the patient died within 5 year

5 Univariate Analysis

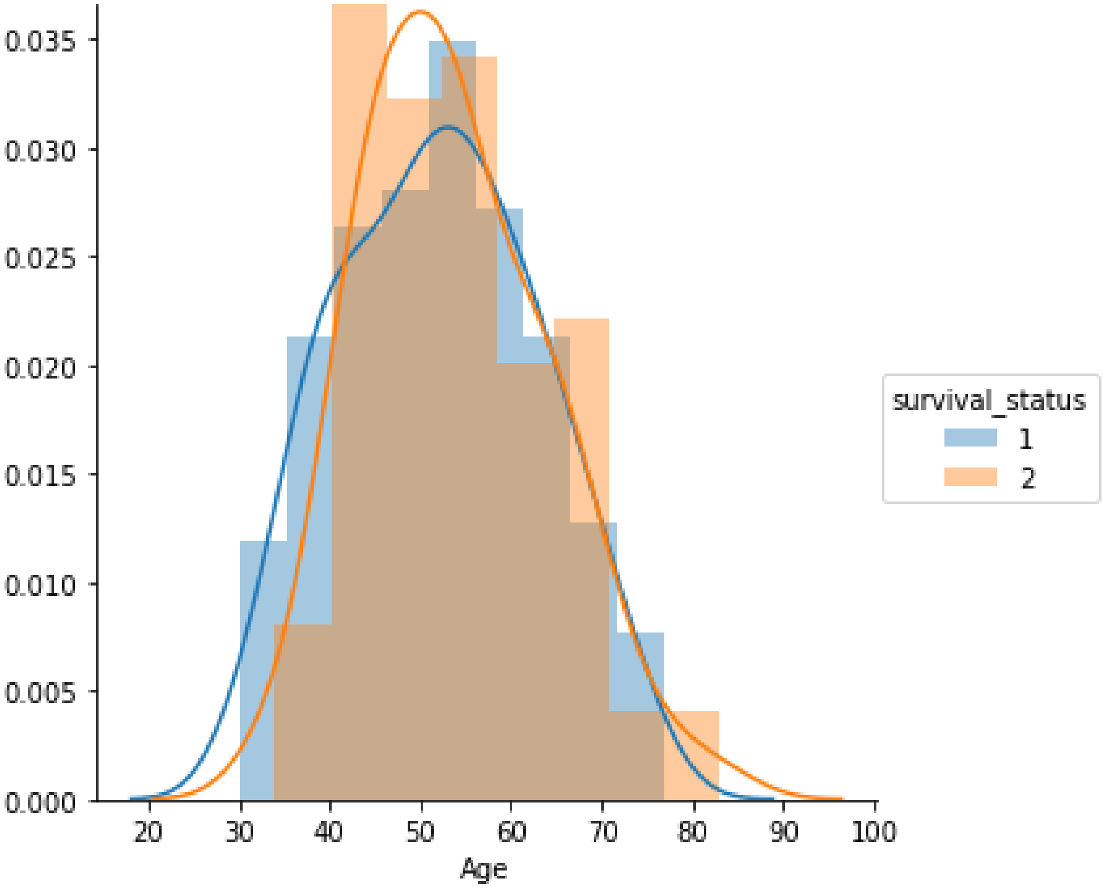
6 Histogram

```
In [6]: se.FacetGrid(hb,hue="survival_status",size=5)\
            .map(se.distplot,"year")\
            .add_legend()
        plt.show()
```



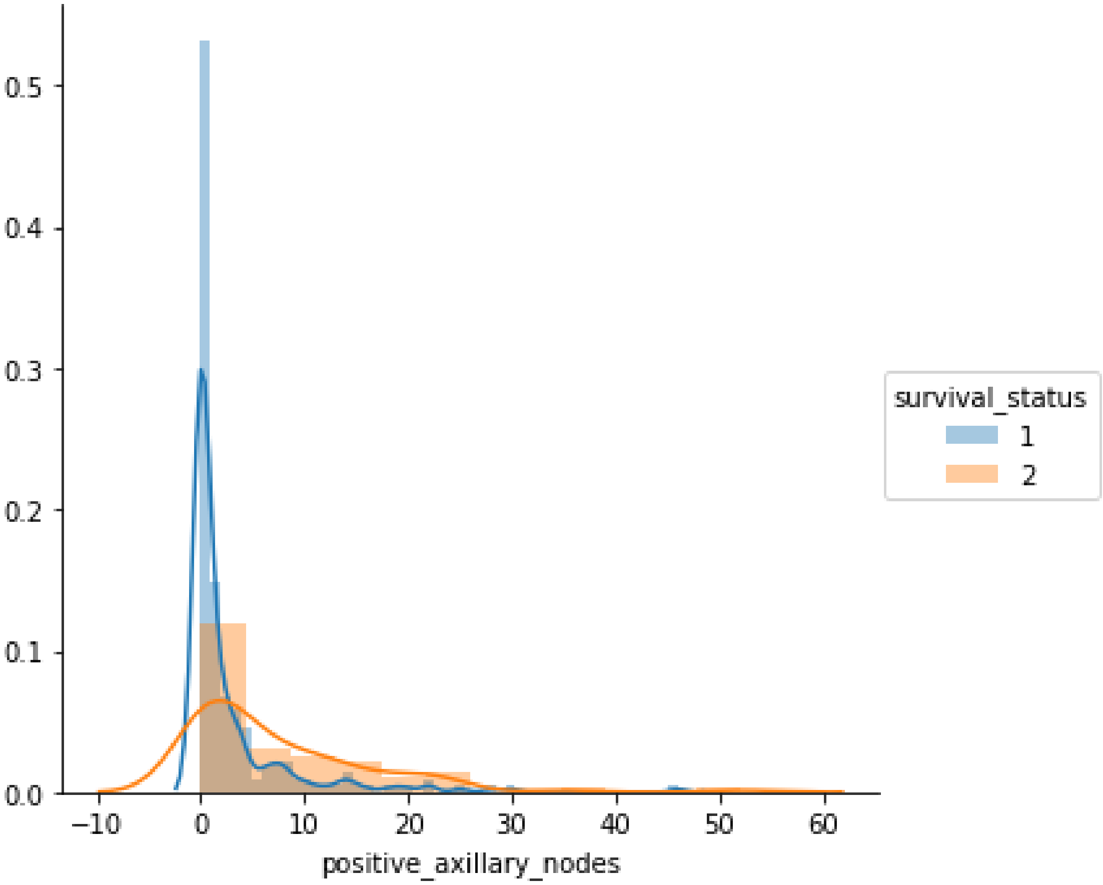
Observation : can't say much from the plot as points are overlapping

```
In [7]: se.FacetGrid(hb,hue="survival_status",size=5)\
            .map(se.distplot,"Age")\
            .add_legend()
        plt.show()
```



Observation : * Patients with age less than 35 and greater than 30 have survived more than 5 years after operation * Patients with age less than 83 and greater than 78 have survived not more than 5 Years after operation * Patients from age 35 to 78 we can't say anything as point are almost overlapping.

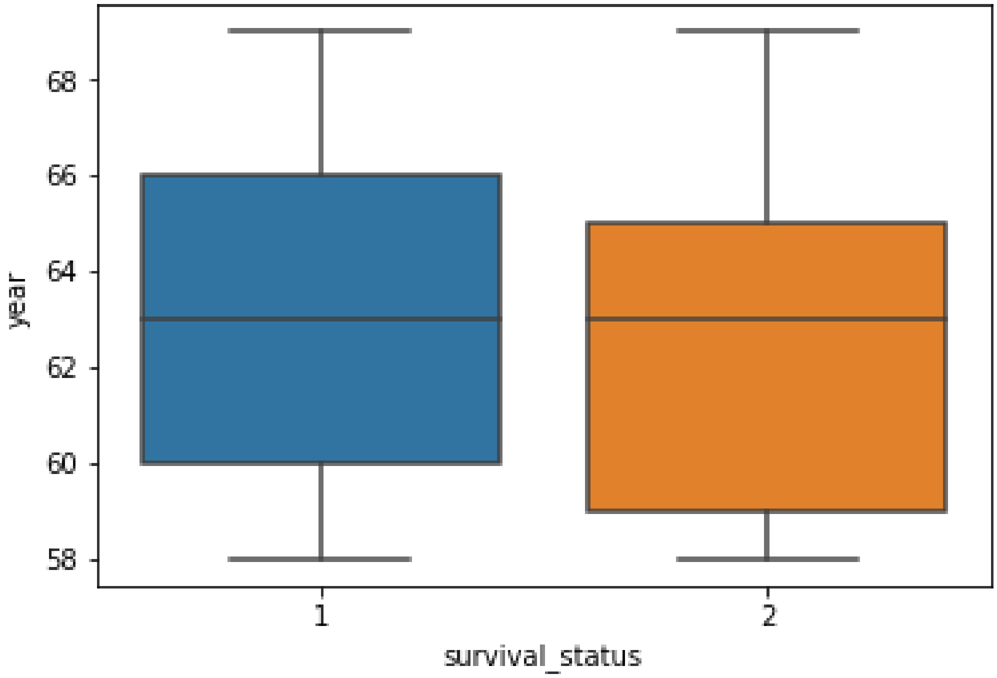
```
In [8]: se.FacetGrid(hb,hue="survival_status",size=5)\
            .map(se.distplot,"positive_axillary_nodes")\
            .add_legend()
        plt.show()
```



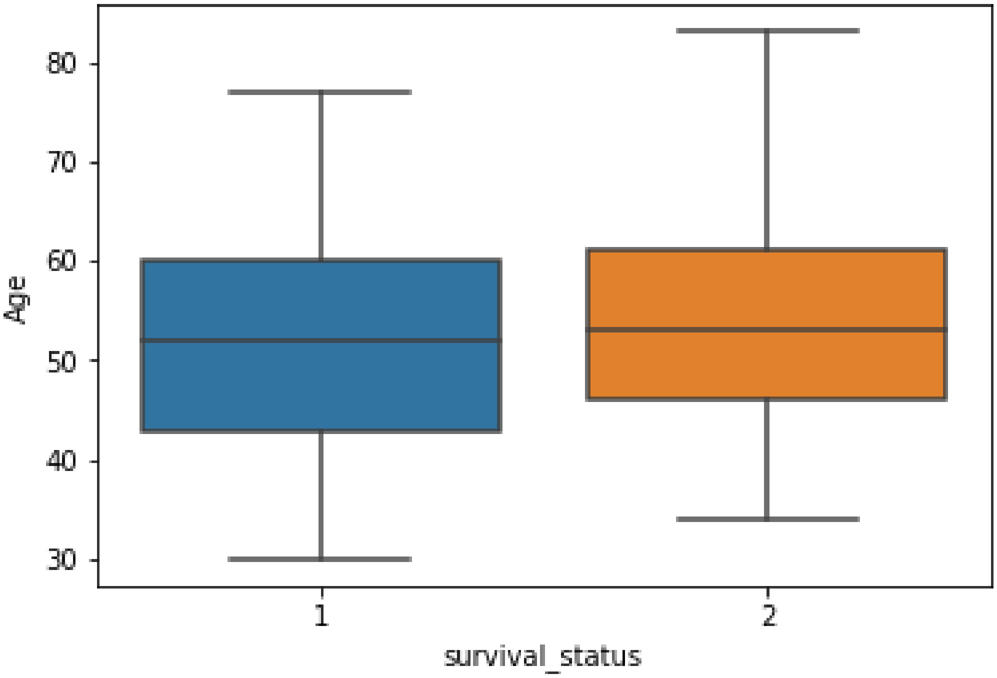
Observation : can't say much from the plot as points are overlapping but one thing we can infer is as the no. of positive auxillary nodes increases the survival status decreases less than 5 years .

7 Box plot and Whiskers

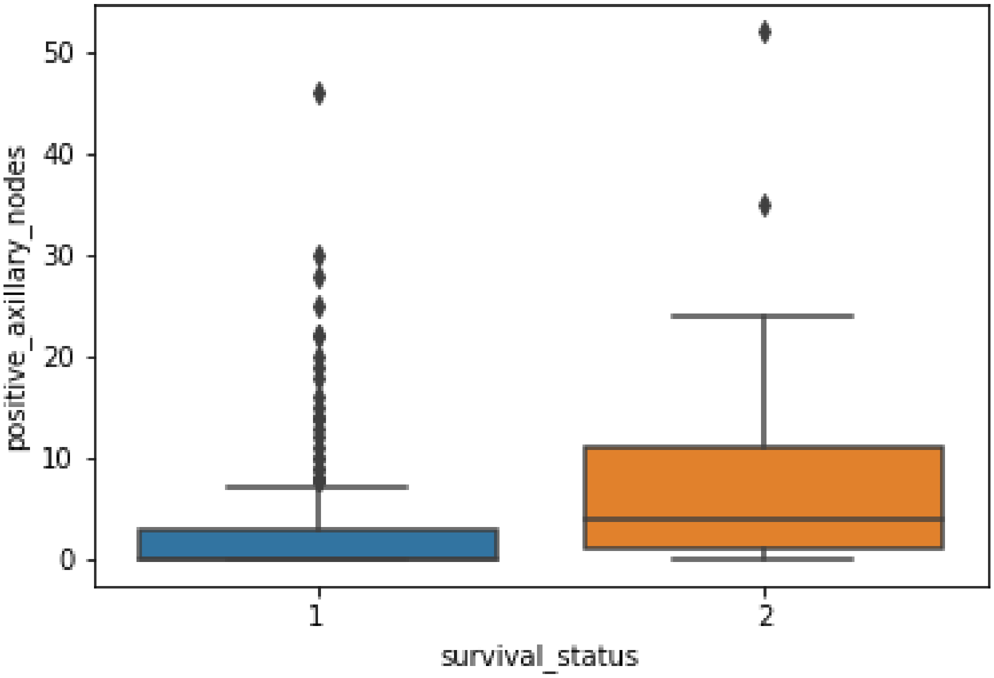
```
In [9]: se.boxplot(x = 'survival_status',y = 'year',data = hb)
        plt.show()
```



```
In [10]: se.boxplot(x = 'survival_status',y = 'Age',data = hb)
         plt.show()
```



```
In [11]: se.boxplot(x = 'survival_status',y = 'positive_axillary_nodes',data = hb)
         plt.show()
```

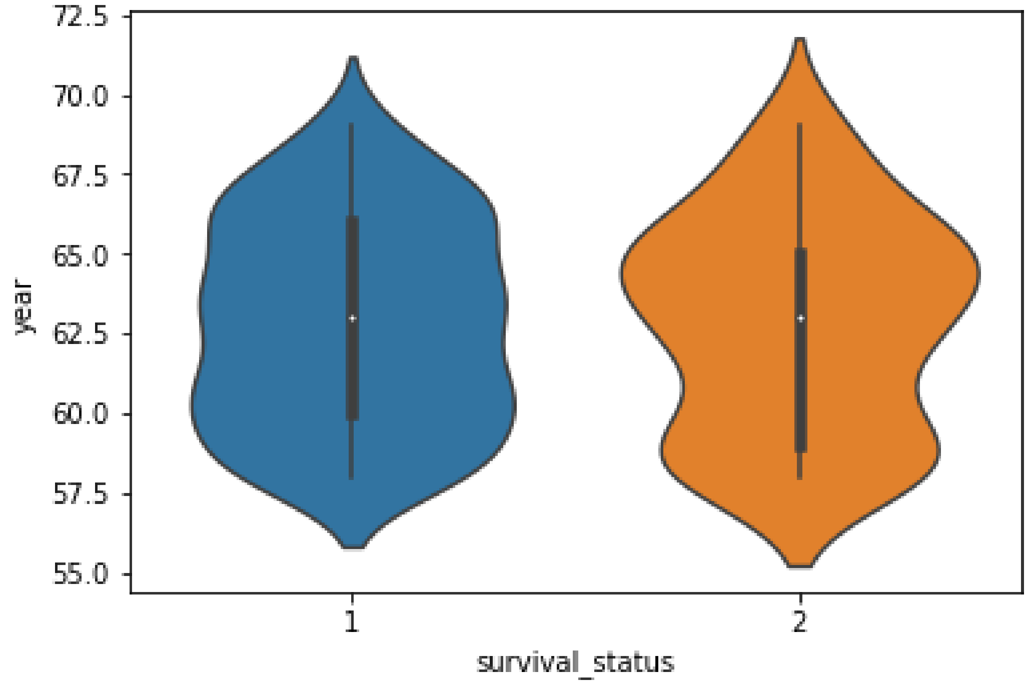


8 Observations

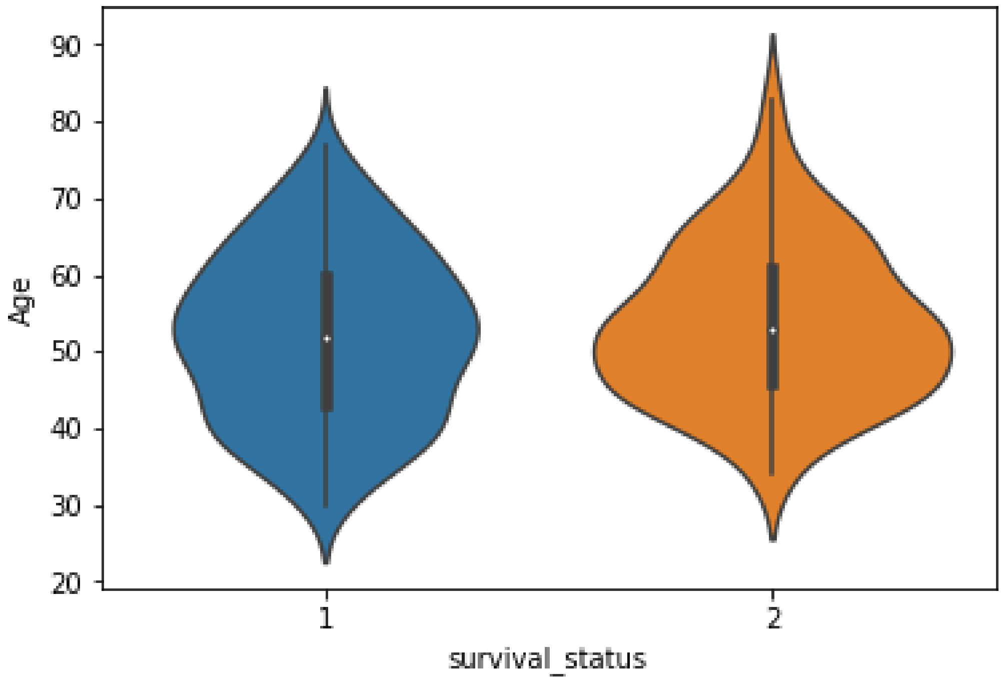
- From the boxplot we can observe that most people who survived cancer have zero positive axillary nodes

9 Violin plots

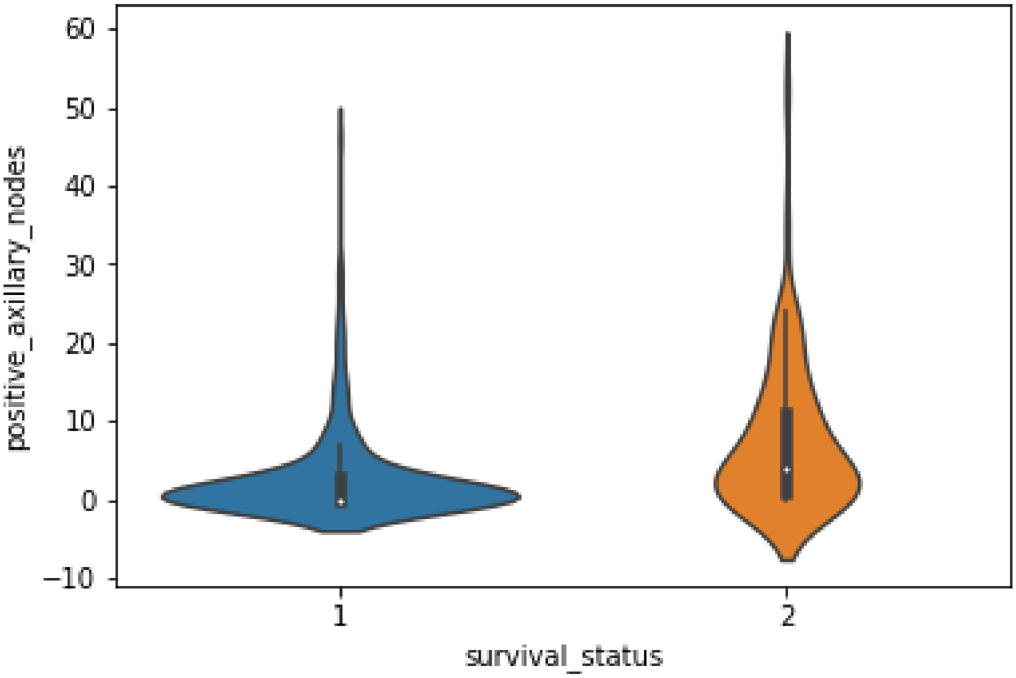
```
In [12]: se.violinplot(x="survival_status", y="year", data=hb, size=8)
         plt.show()
```



```
In [13]: se.violinplot(x="survival_status", y="Age", data=hb, size=8)
         plt.show()
```



```
In [14]: se.violinplot(x="survival_status", y="positive_axillary_nodes", data=hb, size=8)
         plt.show()
```



10 Observation

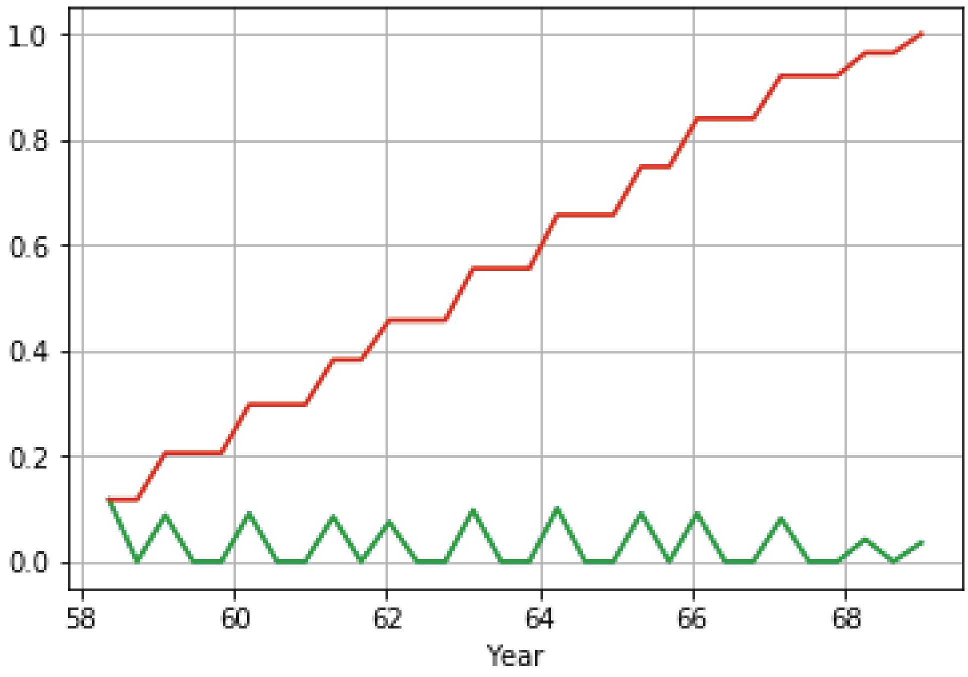
- From the violin plots we can observe that most people who survived cancer have zero positive axillary nodes

11 PDF and CDF

```
In [75]: #pdf cdf of year

         counts,bin_edges = np.histogram(hb['year'],bins = 30, density = True)
         pdf = counts/(sum(counts))
         cdf = np.cumsum(pdf)
         plt.plot(bin_edges[1:],pdf)
         plt.plot(bin_edges[1:],cdf)
         plt.legend()

         counts,bin_edges = np.histogram(hb['year'],bins = 30, density = True)
         pdf = counts/(sum(counts))
         cdf = np.cumsum(pdf)
         plt.plot(bin_edges[1:],pdf)
```

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


        plt.xlabel('Year')
        plt.grid()

        plt.show()
```

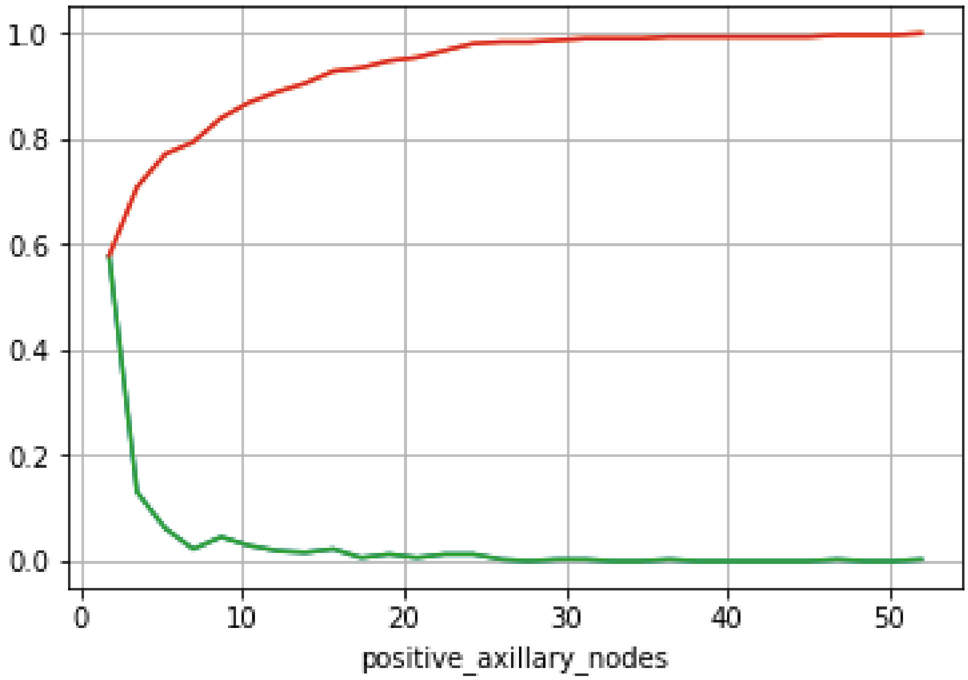
```
C:\Users\sagun\AppData\Local\Continuum\anaconda3\lib\site-packages\matplotlib\axes\_axes.py:545
  warnings.warn("No labelled objects found. "
```



```
In [76]: #pdf cdf of positive_axillary_nodes

         counts,bin_edges = np.histogram(hb['positive_axillary_nodes'],bins = 30, density = Tru
         pdf = counts/(sum(counts))
         cdf = np.cumsum(pdf)
         plt.plot(bin_edges[1:],pdf)
         plt.plot(bin_edges[1:],cdf)
         plt.legend()

         counts,bin_edges = np.histogram(hb['positive_axillary_nodes'],bins = 30, density = Tru
         pdf = counts/(sum(counts))
         cdf = np.cumsum(pdf)
         plt.plot(bin_edges[1:],pdf)
```

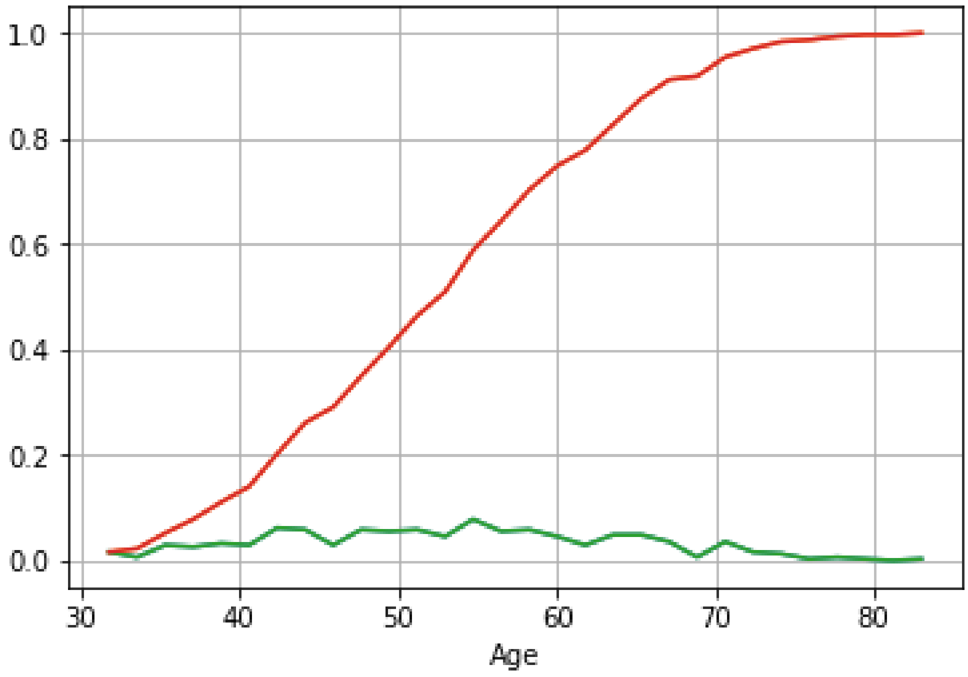
```
        plt.plot(bin_edges[1:],cdf)

        plt.xlabel('positive_axillary_nodes')
        plt.grid()

        plt.show()
```

```
C:\Users\sagun\AppData\Local\Continuum\anaconda3\lib\site-packages\matplotlib\axes\_axes.py:545
  warnings.warn("No labelled objects found. "
```



```
In [77]: #pdf cdf of Age

         counts,bin_edges = np.histogram(hb['Age'],bins = 30, density = True)
         pdf = counts/(sum(counts))
         cdf = np.cumsum(pdf)
         plt.plot(bin_edges[1:],pdf)
         plt.plot(bin_edges[1:],cdf)
         plt.legend()

         counts,bin_edges = np.histogram(hb['Age'],bins = 30, density = True)
         pdf = counts/(sum(counts))
         cdf = np.cumsum(pdf)
         plt.plot(bin_edges[1:],pdf)
         plt.plot(bin_edges[1:],cdf)
         plt.xlabel('Age')
         plt.grid()

         plt.show()
```

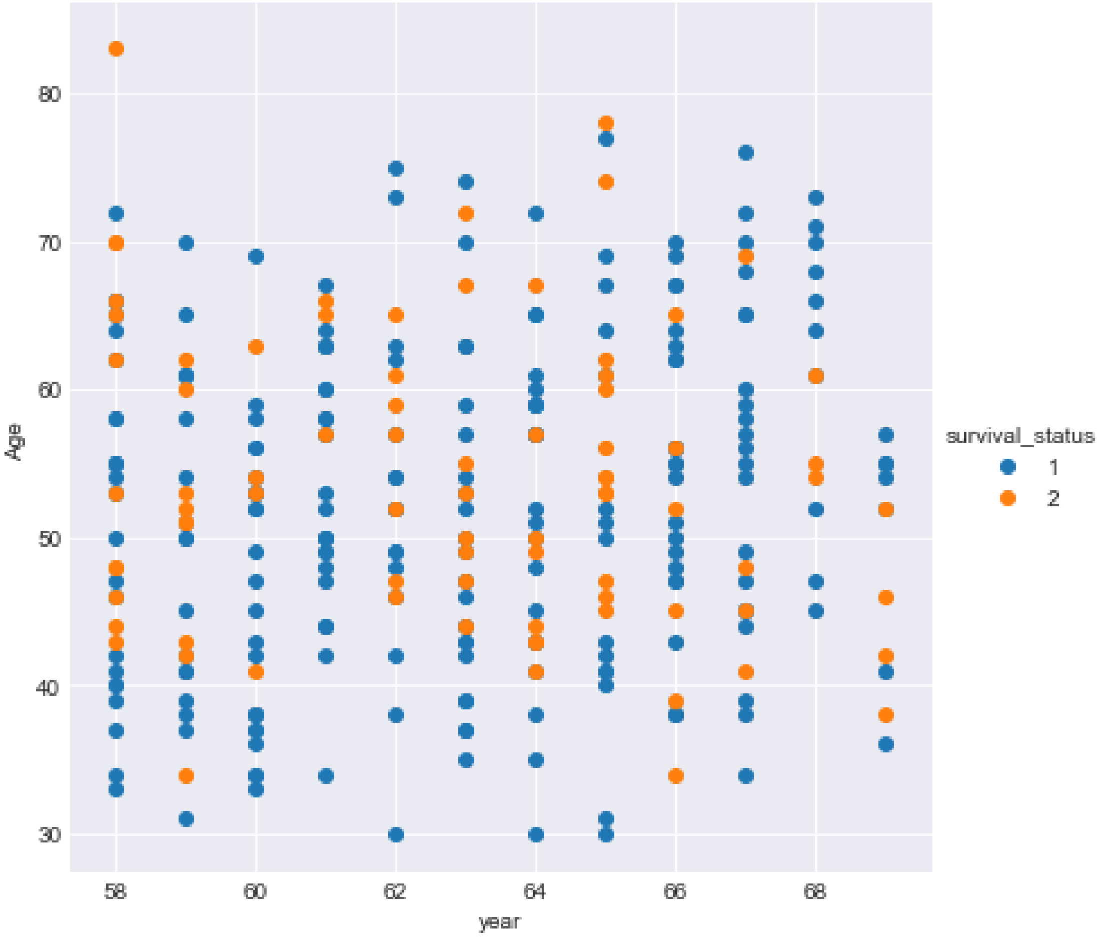
```
C:\Users\sagun\AppData\Local\Continuum\anaconda3\lib\site-packages\matplotlib\axes\_axes.py:545
  warnings.warn("No labelled objects found. "
```



12 Bivariate analysis

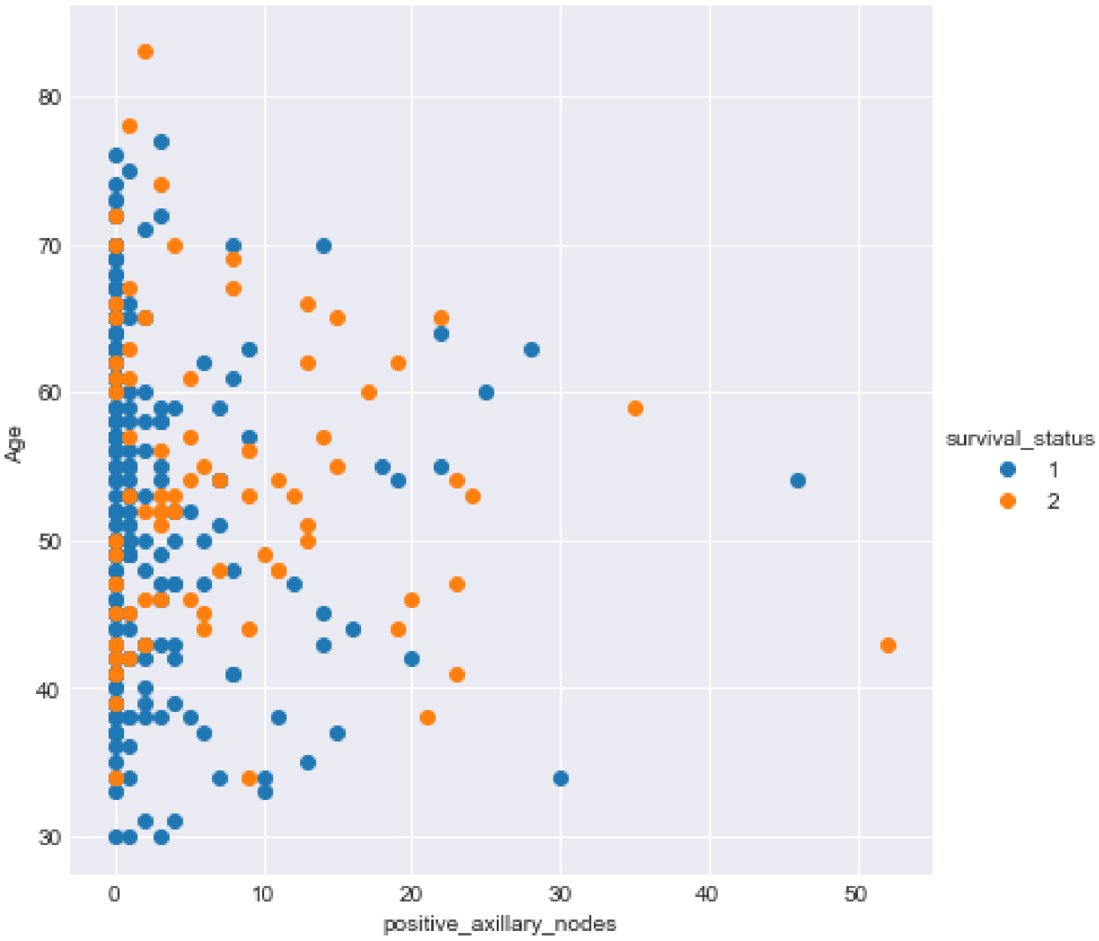
13 2-D Scatter Plot

```
In [16]: se.set_style("darkgrid");
         se.FacetGrid(hb,hue='survival_status',size=6)\
             .map(plt.scatter,"year","Age")\
             .add_legend();
         plt.show()
```



Observation : can't say much from the plot as points overlapping

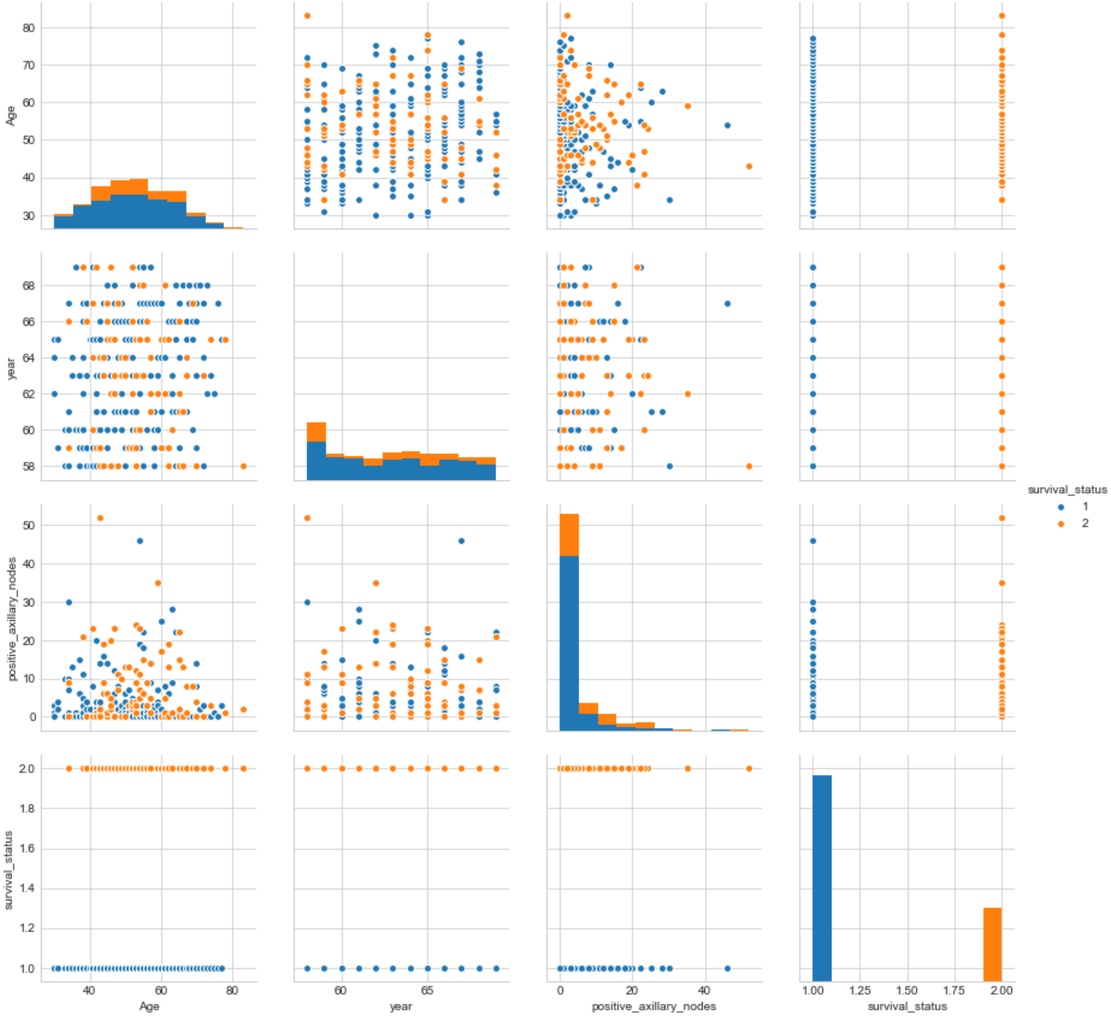
```
In [19]: se.set_style("darkgrid");
         se.FacetGrid(hb,hue='survival_status',size=6)\
             .map(plt.scatter,"positive_axillary_nodes","Age")\
             .add_legend();
         plt.show()
```



Observation : can't say much from the plot as points overlapping

14 Pair-Plot

```
In [13]: plt.close();
         se.set_style("whitegrid");
         se.pairplot(hb,hue="survival_status",size=3)
         plt.show()
```



15 Observations

- Positive_axillary_nodes is a useful feature to identify the survival_status of cancer patients
- Age and Year of operation have overlapping curves so we can't have a suitable observation that can classify survival_status

16 Mean

```
In [80]: #hb is the name of the data frame
         less_five = hb[hb['survival_status']==2]
         more_five = hb[hb['survival_status']==1]

In [73]: print(np.mean(more_five))
```

```
Age                        52.017778
year                       62.862222
positive_axillary_nodes     2.791111
survival_status             1.000000
dtype: float64
```

```
In [74]: print(np.mean(less_five))
```

```
Age                        53.679012
year                       62.827160
positive_axillary_nodes     7.456790
survival_status             2.000000
dtype: float64
```

Observation * Mean age of patients who survived more than 5 years is 52 years and who didn't survive is 54 years * Those having more than 3 positive_axillary_nodes they have not survived more than 5 years * Those having less than 3 positive_axillary_nodes they have survived more than 5 years after the operation

17 Final Conclusion

- Those having more than 3 positive_axillary_nodes they have not survived more than 5 years
- Those having less than 3 positive_axillary_nodes they have survived more than 5 years after the operation
- Positive_axillary_nodes is a useful feature to identify the survival_status of cancer patients
- Age and Year of operation have overlapping curves so we can't classify patients for their survival_status using age