

Haberman's Survival Exercise

March 10, 2018

1 Exercise

1.0.1 Haberman's Survival

Survival of patients who had undergone surgery for breast cancer

```
In [2]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [3]: patients = pd.read_csv('haberman.csv')
patients.shape
```

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

Observation: There are total 305 data points There are 3 Features/dimensions

Columns Information:

1)Age of patient at time of operation (numerical) 2)Patient's year of operation (year - 1900, numerical) 3)Number of positive axillary nodes detected (numerical) 4)Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the patient died within 5 year

```
In [4]: patients.head()
```

```
Out[4]:
```

	Age	Year	No.of Pos Axillary nodes	Survival 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

```
In [5]: patients.describe()
```

```
Out[5]:
```

	Age	Year	No.of Pos Axillary nodes	Survival 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

Observation: The Average age is 30

```
In [6]: # getting survival status count of 2 groups
```

```
patients.get('Survival Status').value_counts()
```

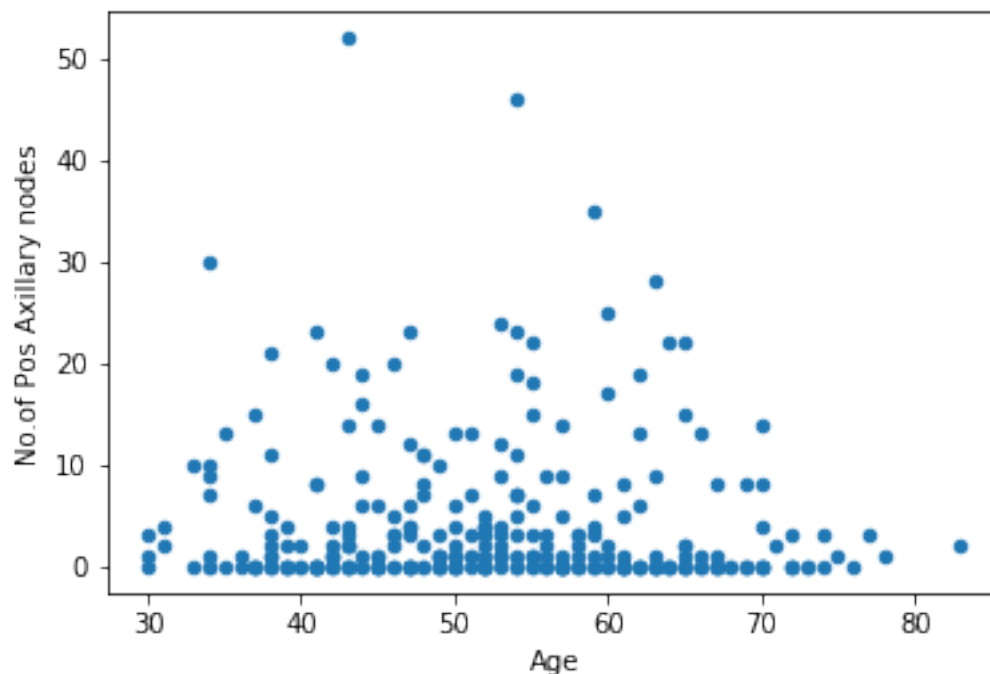
```
Out[6]: 1    225  
        2     81  
        Name: Survival Status, dtype: int64
```

Observation: Survived: 225 out of 305 Dead with in 5 years: 81 out of 305
Its Imbalanced Data Set.

1.1 BiVariant Analysis

1.1.1 2D Scatter Plot

```
In [7]: patients.plot(y='No.of Pos Axillary nodes', x='Age',kind='scatter');  
plt.show()
```

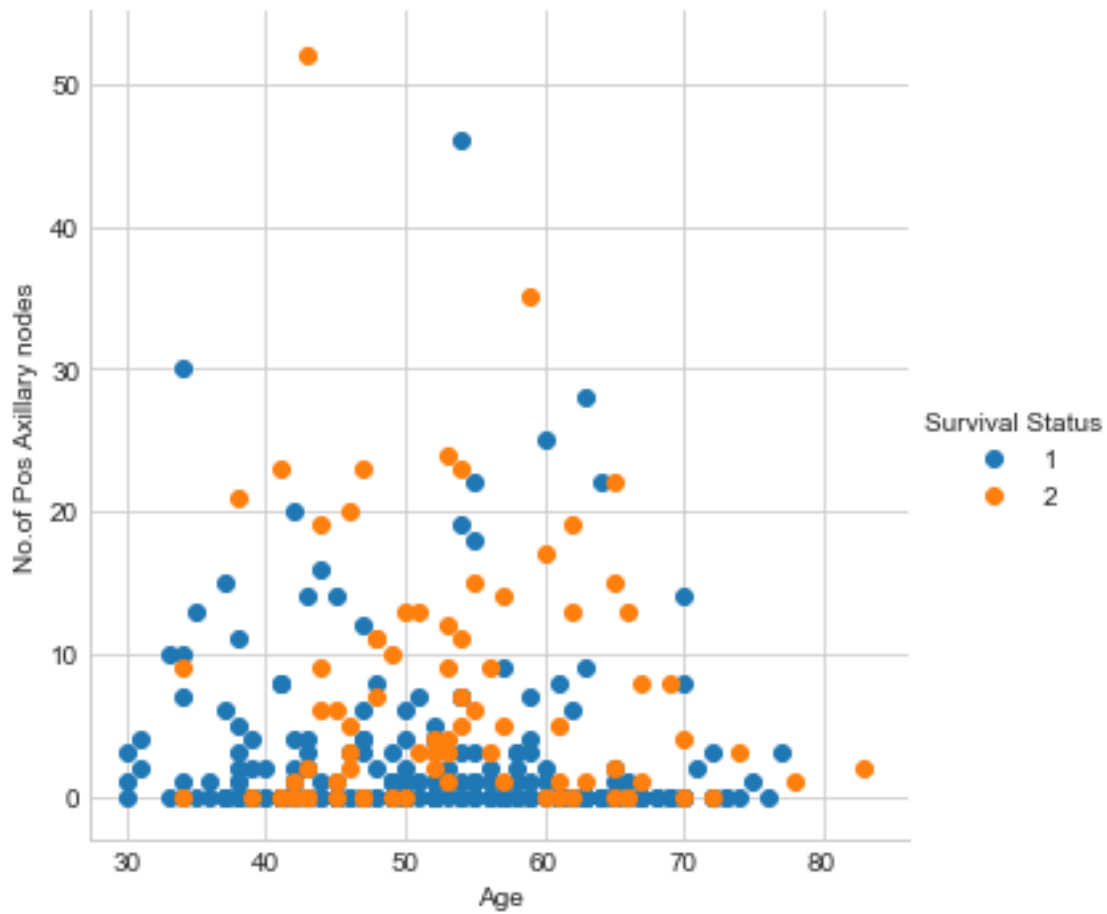


Observation:
For Most of the Patients The No.of axillary nodes is '0'

```
In [8]: sns.set_style("whitegrid")
```

```
sns.FacetGrid(patients,hue='Survival Status',size=5) \
```

```
.map(plt.scatter, 'Age', 'No. of Pos Axillary nodes') \
.add_legend()
plt.show()
```

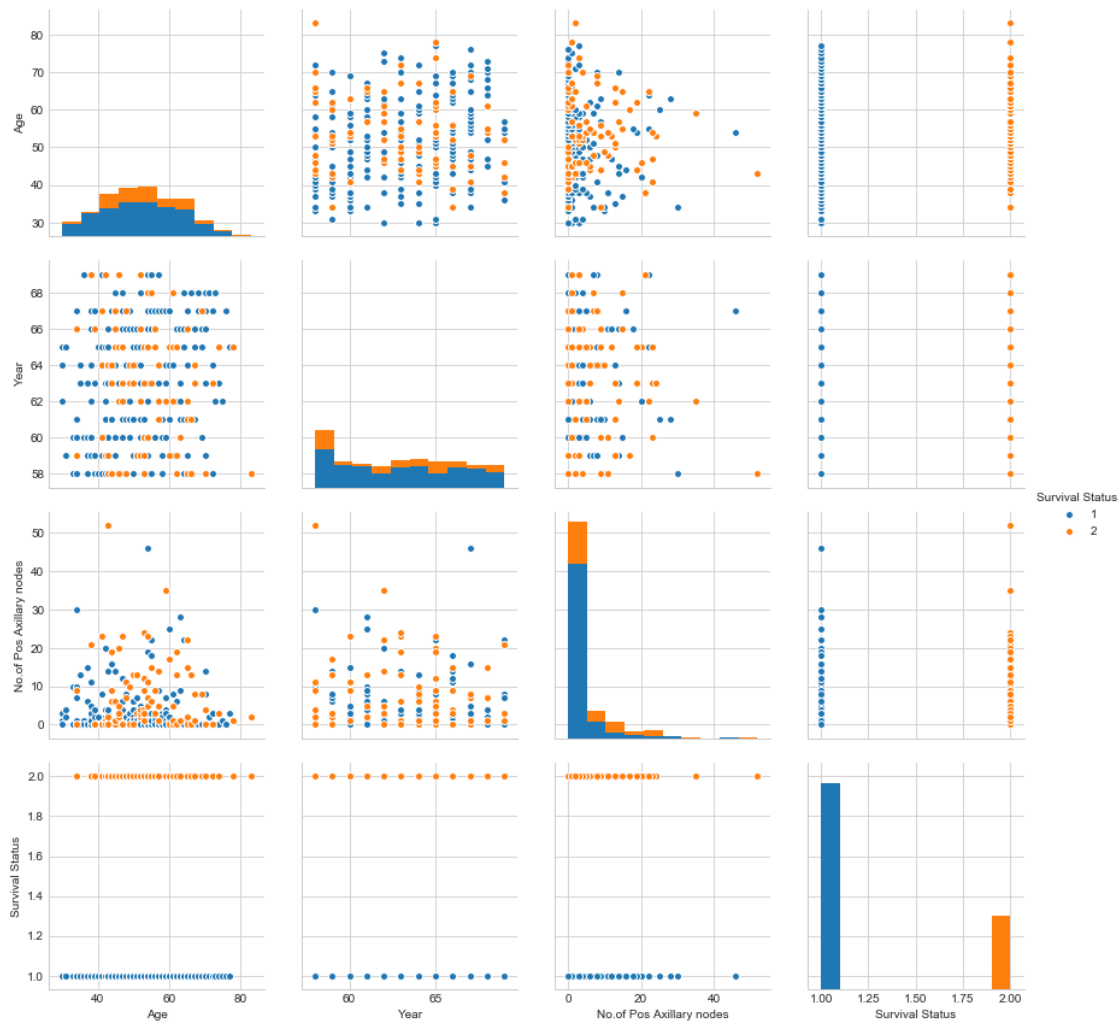


Observation: We cant classify through this features with this type of plot.

1.1.2 PairPlot

```
In [9]: sns.set_style("whitegrid");

sns.pairplot(patients,hue='Survival Status',size=3)
plt.show()
```

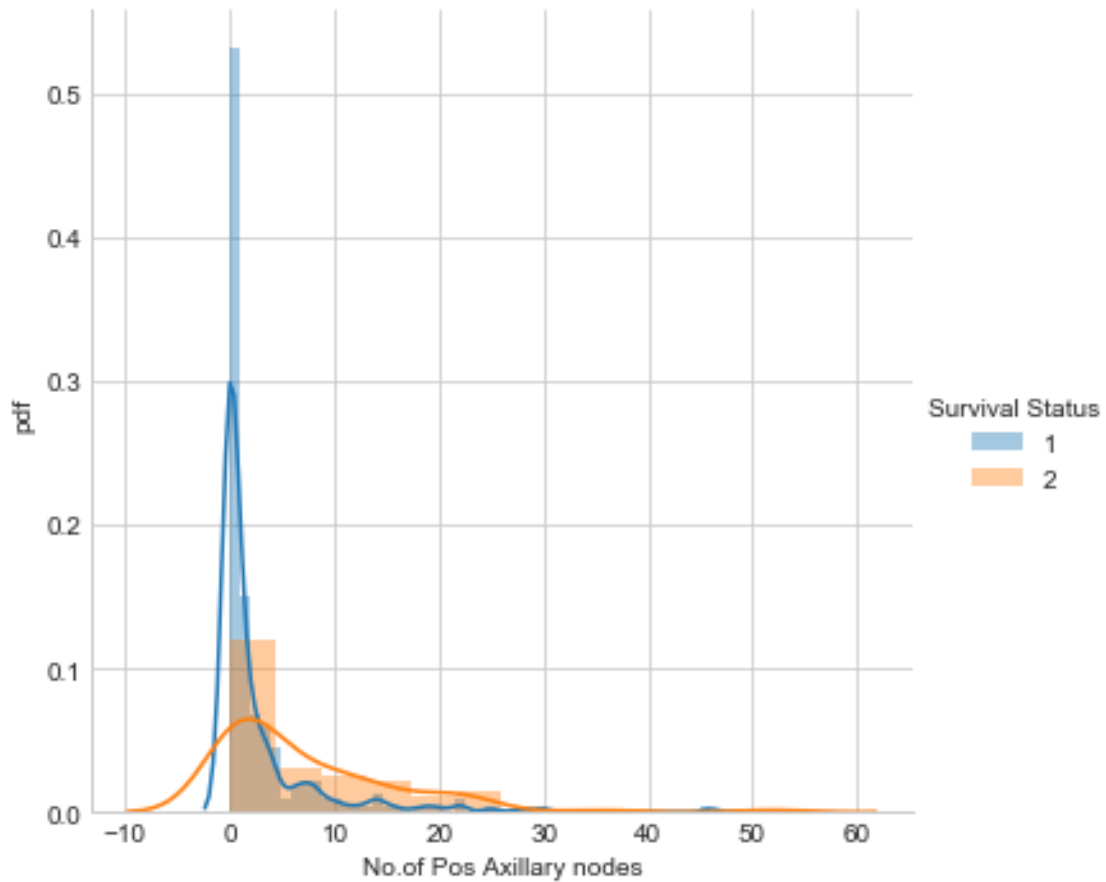


Note: Without Considering Survival statues column there are 3 cols-> 3C2 plots ->3 plots
With this 3 plots its difficult to classify the survival status.

1.2 Univariant Analysis

1.2.1 PDF

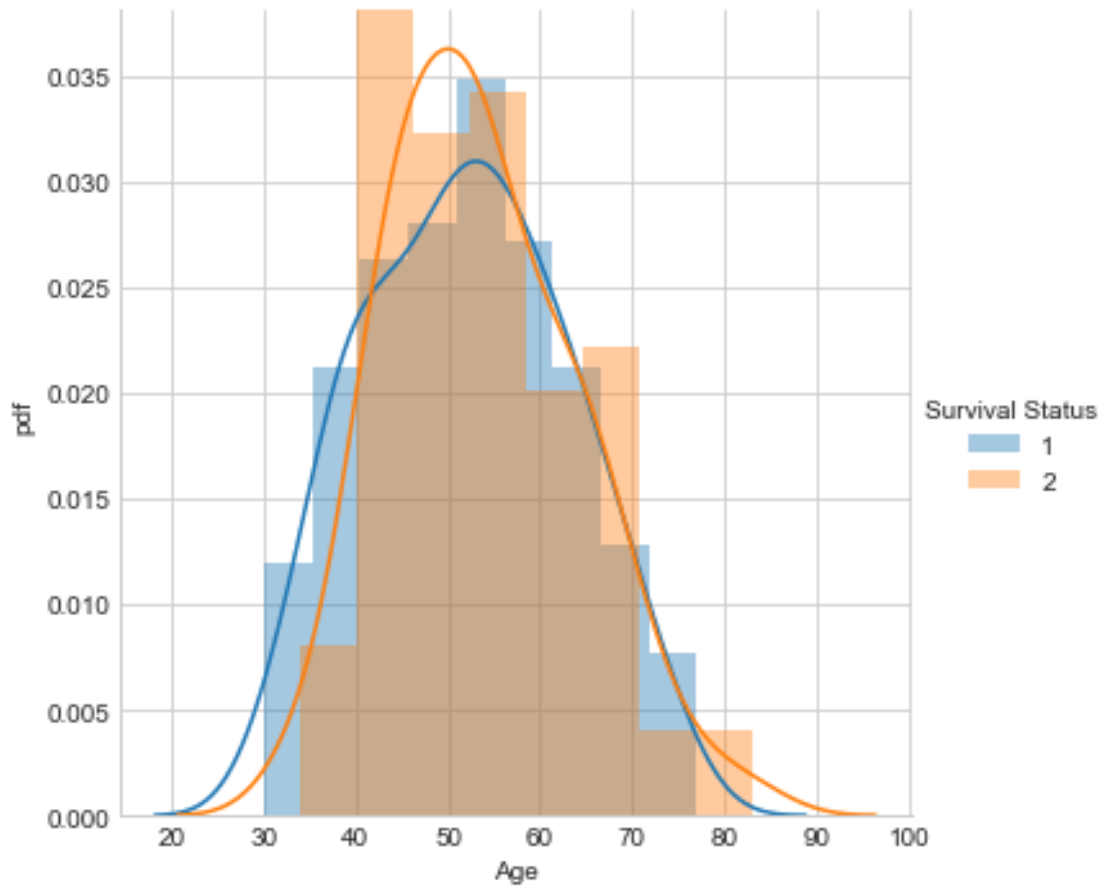
```
In [10]: sns.FacetGrid(patients, hue="Survival Status", size=5) \
        .map(sns.distplot, "No. of Pos Axillary nodes") \
        .add_legend();
plt.ylabel('pdf')
plt.show();
```



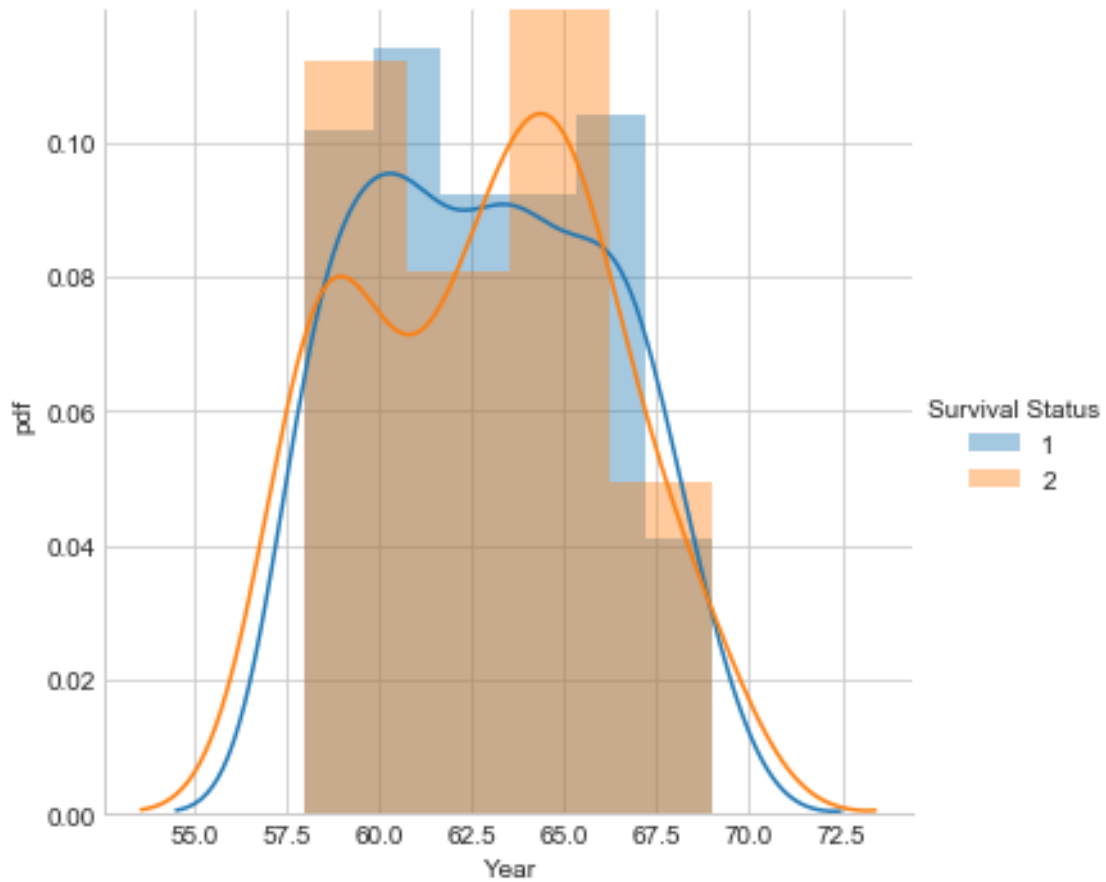
Observation: Spread is more for status 1 than status 2

More no. of patients of status 1 has less no. of Axillary Nodes That may indicate an inverse relation between Axillary nodes and Survival Status

```
In [11]: sns.FacetGrid(patients, hue="Survival Status", size=5) \
        .map(sns.distplot, "Age") \
        .add_legend();
plt.ylabel('pdf')
plt.show();
```



```
In [12]: sns.FacetGrid(patients, hue="Survival Status", size=5) \
        .map(sns.distplot, "Year") \
        .add_legend();
plt.ylabel('pdf')
plt.show();
```



```
In [13]: status_1= patients[patients['Survival Status']==1]
status_1.head()
```

```
Out[13]:
```

	Age	Year	No.of Pos Axillary nodes	Survival 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

```
In [14]: status_2= patients[patients['Survival Status']==2]
status_2.head()
```

```
Out[14]:
```

	Age	Year	No.of Pos Axillary nodes	Survival Status
7	34	59	0	2
8	34	66	9	2
24	38	69	21	2
34	39	66	0	2
43	41	60	23	2

```
In [15]: status_1.describe()
```

```
Out[15]:
```

	Age	Year	No.of Pos Axillary nodes	Survival Status
count	225.000000	225.000000	225.000000	225.0
mean	52.017778	62.862222	2.791111	1.0
std	11.012154	3.222915	5.870318	0.0
min	30.000000	58.000000	0.000000	1.0
25%	43.000000	60.000000	0.000000	1.0
50%	52.000000	63.000000	0.000000	1.0
75%	60.000000	66.000000	3.000000	1.0
max	77.000000	69.000000	46.000000	1.0

```
In [16]: status_2.describe()
```

```
Out[16]:
```

	Age	Year	No.of Pos Axillary nodes	Survival Status
count	81.000000	81.000000	81.000000	81.0
mean	53.679012	62.827160	7.456790	2.0
std	10.167137	3.342118	9.185654	0.0
min	34.000000	58.000000	0.000000	2.0
25%	46.000000	59.000000	1.000000	2.0
50%	53.000000	63.000000	4.000000	2.0
75%	61.000000	65.000000	11.000000	2.0
max	83.000000	69.000000	52.000000	2.0

Observation: Status 1: Mean of no.of axillary nodes -> 2.8~ std -> 5.9~ Status 2: Mean of no.of axillary nodes -> 7.5~ std -> 9.1~

That Implies More no.of Axillary nodes patients may tend to die with in 5 years of operation

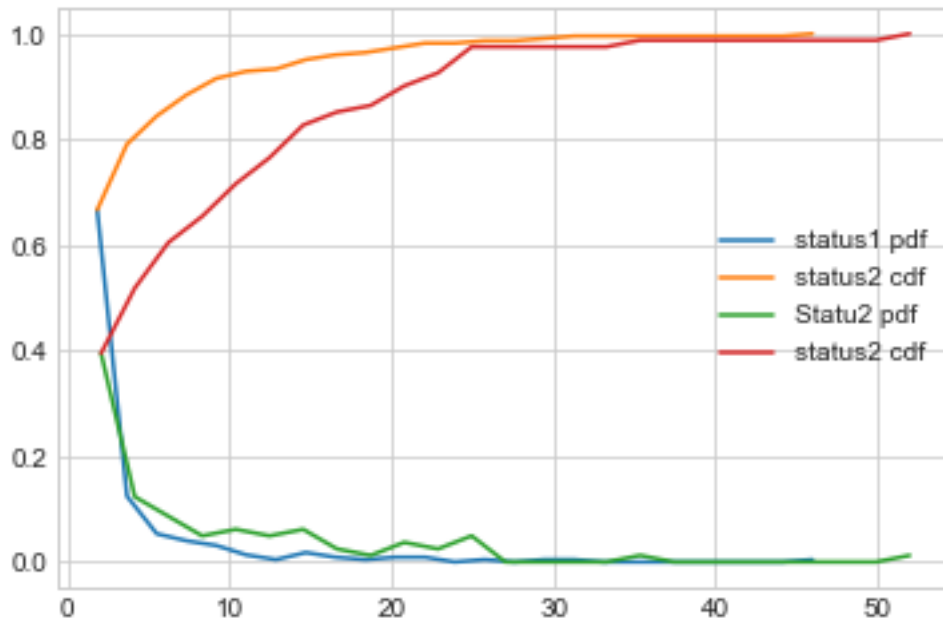
1.2.2 PDF CDF

```
In [17]: counts, bin_edges = np.histogram(status_1['No.of Pos Axillary nodes'], bins=25,  
                                           density = True)
```

```
pdf = counts/(sum(counts))  
cdf = np.cumsum(pdf)  
plt.plot(bin_edges[1:],pdf,label='status1 pdf');  
plt.plot(bin_edges[1:], cdf,label='status2 cdf')
```

```
counts, bin_edges = np.histogram(status_2['No.of Pos Axillary nodes'], bins=25,  
                                 density = True)
```

```
pdf = counts/(sum(counts))  
cdf = np.cumsum(pdf)  
plt.plot(bin_edges[1:],pdf,label='Statu2 pdf');  
plt.plot(bin_edges[1:], cdf,label='status2 cdf')  
plt.legend()  
plt.show()
```

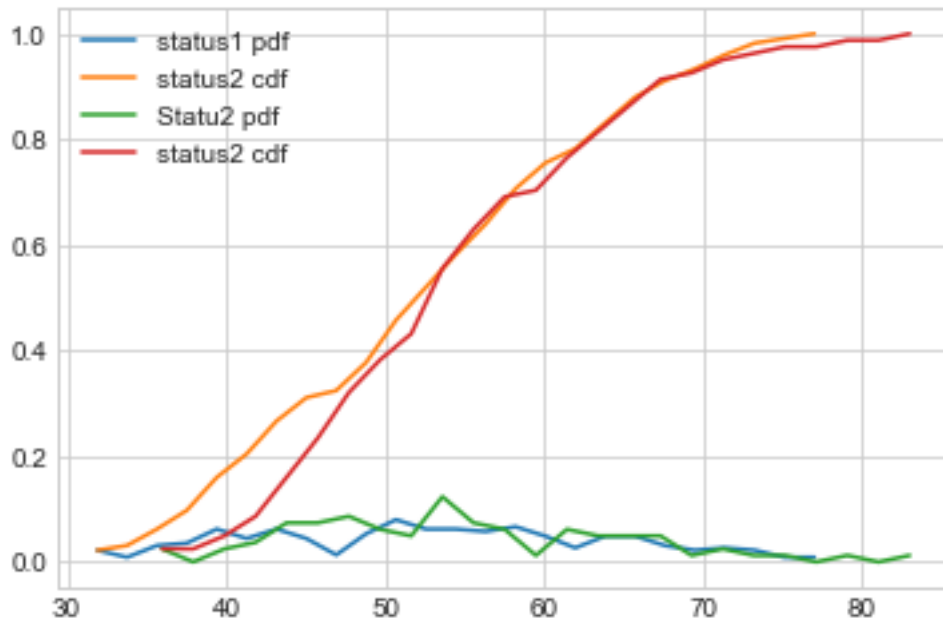



```
In [18]: counts, bin_edges = np.histogram(status_1['Age'], bins=25,
                                           density = True)

pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label='status1 pdf');
plt.plot(bin_edges[1:], cdf,label='status2 cdf')

counts, bin_edges = np.histogram(status_2['Age'], bins=25,
                                 density = True)

pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label='Statu2 pdf');
plt.plot(bin_edges[1:], cdf,label='status2 cdf')
plt.legend()
plt.show()
```

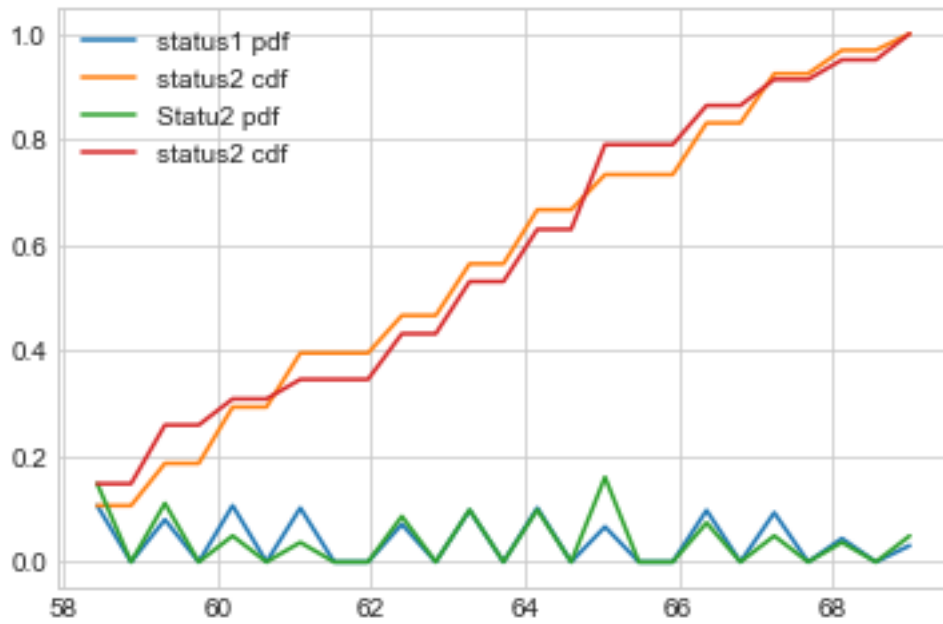


```
In [19]: counts, bin_edges = np.histogram(status_1['Year'], bins=25,
                                           density = True)

pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label='status1 pdf');
plt.plot(bin_edges[1:], cdf,label='status2 cdf')

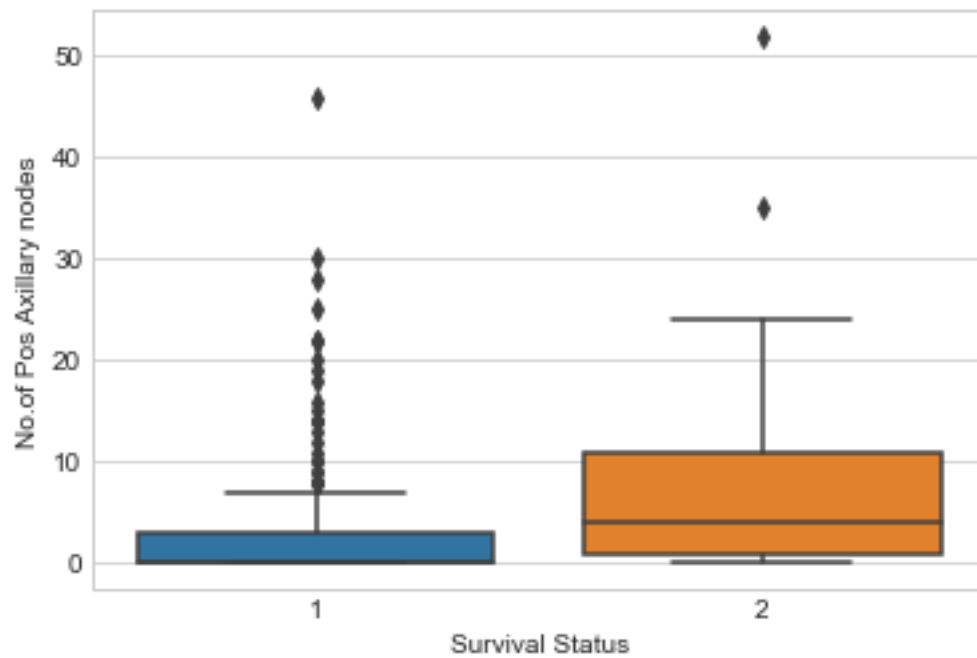
counts, bin_edges = np.histogram(status_2['Year'], bins=25,
                                 density = True)

pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label='Statu2 pdf');
plt.plot(bin_edges[1:], cdf,label='status2 cdf')
plt.legend()
plt.show()
```



1.2.3 Box Plot

In [20]: `sns.boxplot(x='Survival Status',y='No.of Pos Axillary nodes', data=patients)
plt.show()`

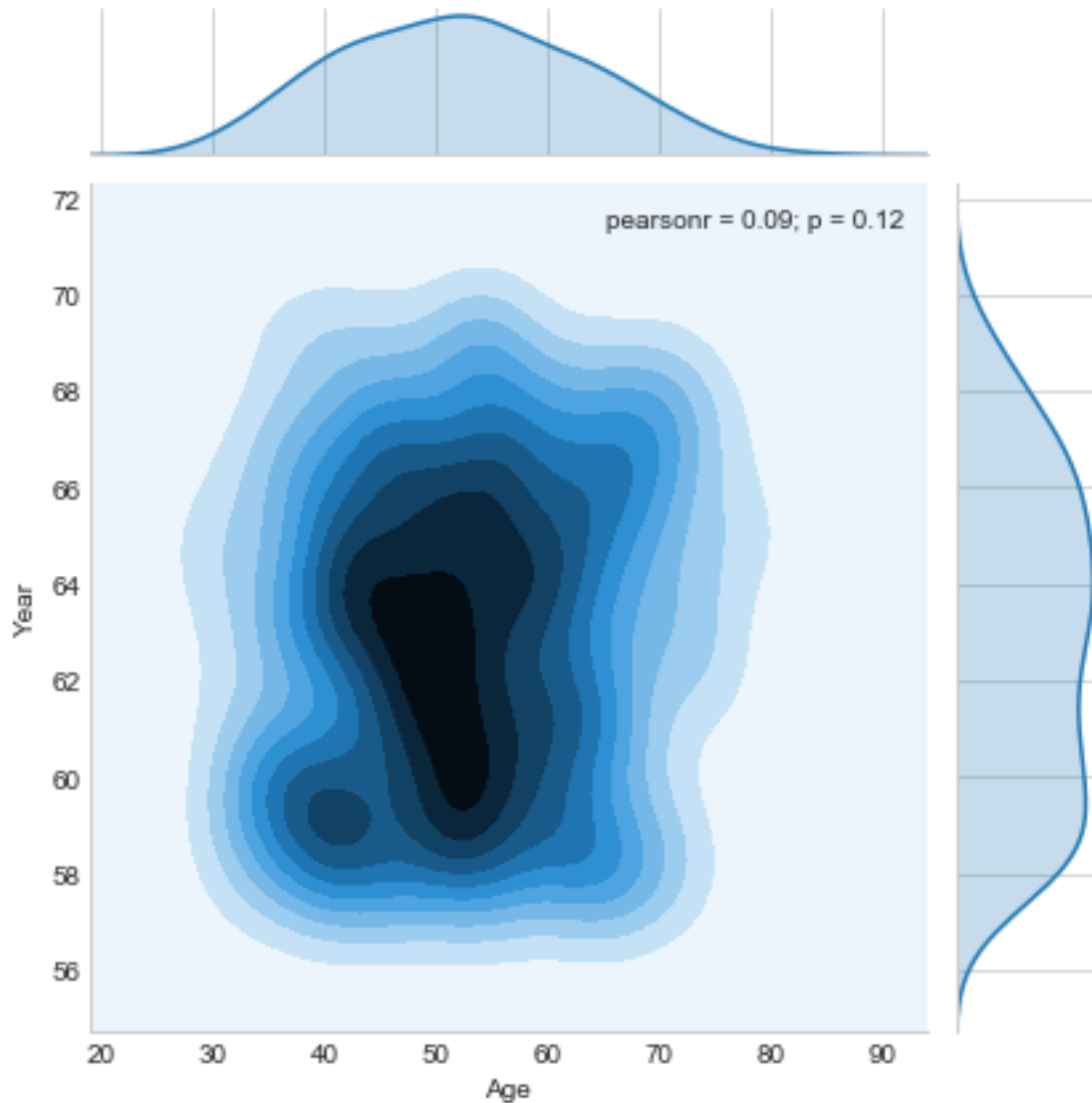


Observation:

Status1 -> About 50 percentile people has Axillary nodes 0 and 75 percentile have about 3~ axillary nodes. status2 -> About 50 percentile people has Axillary nodes more than 4~ and 25 percentile patients have 1-3 nodes.

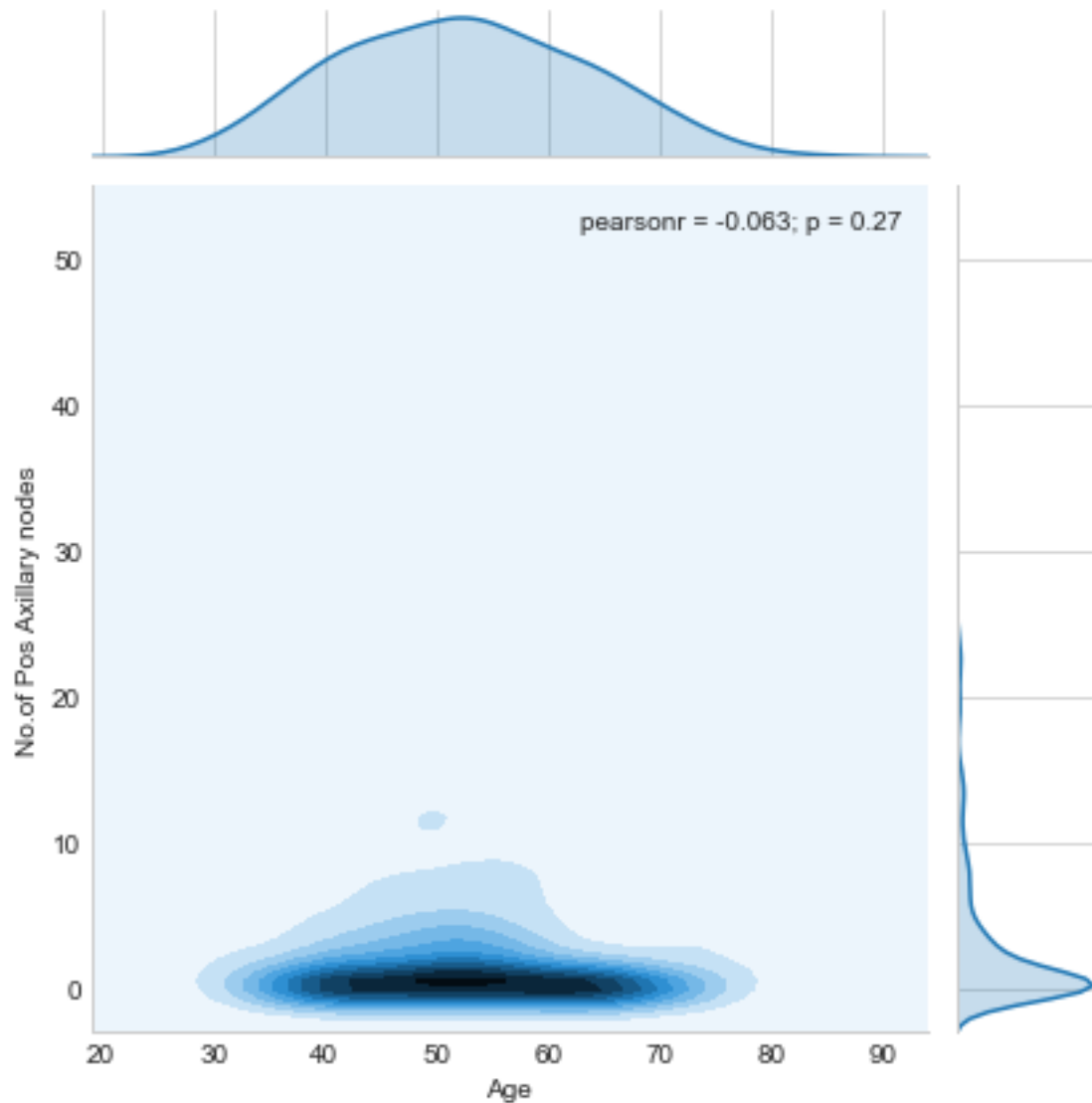
1.2.4 Violin Plot

```
In [21]: sns.violinplot(x='Survival Status',y='No.of Pos Axillary nodes', data=patients,size=8,  
plt.show())
```



1.3 MultiVariant Analysis

```
In [22]: sns.jointplot(x= 'Age',kind = 'kde', y='Year', data = patients)  
plt.show()
```



There are more number of patients undergone operation during the year 1960 - 1964 period and between ages 45 - 55

```
In [23]: sns.jointplot(x= 'Age',kind = 'kde', y='No.of Pos Axillary nodes', data = patients)
plt.show()
```

