

1HabermanAnalysis

October 10, 2018

```
In [1]: import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

import warnings; warnings.simplefilter('ignore')

In [2]: datasets_path = '/home/monodeepdas112/container-vrts/AppliedAICourse/Datasets/'
dataset = 'haberman.csv'
haberman = pd.read_csv(datasets_path+dataset)
```

0.1 Description of the dataset

```
In [3]: #Converting to proper dataset column names
haberman.columns = pd.Series(['Age', 'Op_Year', 'Num_Ax_nodes', 'Survival_Status'])

In [4]: #Shape
print('Dataset Shape : ', haberman.shape)

#Description of dataset
print('\nDescription :\n', haberman.describe())

#Columns in the dataset
print('\n\nColumns in the dataset : ', haberman.columns)
```

Dataset Shape : (305, 4)

Description :

	Age	Op_Year	Num_Ax_nodes	Survival_Status
count	305.000000	305.000000	305.000000	305.000000
mean	52.531148	62.849180	4.036066	1.265574
std	10.744024	3.254078	7.199370	0.442364
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%	61.000000	66.000000	4.000000	2.000000
max	83.000000	69.000000	52.000000	2.000000

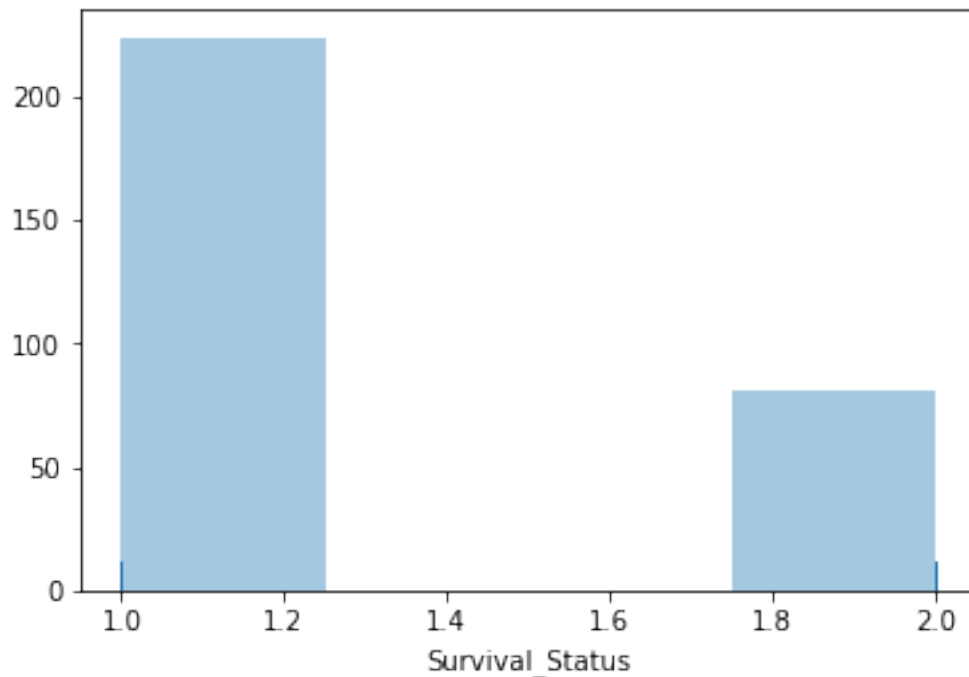
Columns in the dataset : `Index(['Age', 'Op_Year', 'Num_Ax_nodes', 'Survival_Status'], dtype='object')`

0.1.1 Conclusions from dataset description

1. Total number of samples is 305
2. There are 4 Attributes Age, Operation Year, Number of Axil nodes found, Survival Status
3. Relevant statistical indormation can be found out in the cell above

Q. How many people lived more than 5 years and how many less than that ?

```
In [5]: sns.distplot(haberman.Survival_Status, kde=False, rug=True);
```



Approx 225 people who lived more than 5 years after their operation and approx 75 people who lived less than 5 years

Q. How many Number of Axil nodes were found for the people who lived more than 5 years ?

```
In [6]: long_lived=haberman.loc[haberman['Survival_Status'] == 1]
        print('Number of people who lived for more than 5 years : \n\n', long_lived.describe())
```

```

short_lived=haberman.loc[haberman['Survival_Status']==2]
print('\n\nNumber of people who lived for less than 5 years : \n', short_lived.describe())

```

Number of people who lived for more than 5 years :

	Age	Op_Year	Num_Ax_nodes	Survival_Status
count	224.000000	224.000000	224.000000	224.0
mean	52.116071	62.857143	2.799107	1.0
std	10.937446	3.229231	5.882237	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

Number of people who lived for less than 5 years :

	Age	Op_Year	Num_Ax_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

Mean Number of Positive Axil Nodes : 7.456 for the people who could not live for more than 5 years

Mean Number of Positive Axil Nodes : 2.799 for the people who could live for more than 5 years

Gives us a clear picture of the spread of the disease and thus (intuitively) can be used as a important reason for the early death of the people as in their case the disease spread a lot more.

```

In [11]: #Finding out the medians
print('Median Number of Positive Axil Nodes found for class 1 : ', np.median(long_lived['Num_Ax_nodes']))
print('Median Number of Positive Axil Nodes found for class 2 : ', np.median(short_lived['Num_Ax_nodes']))

```

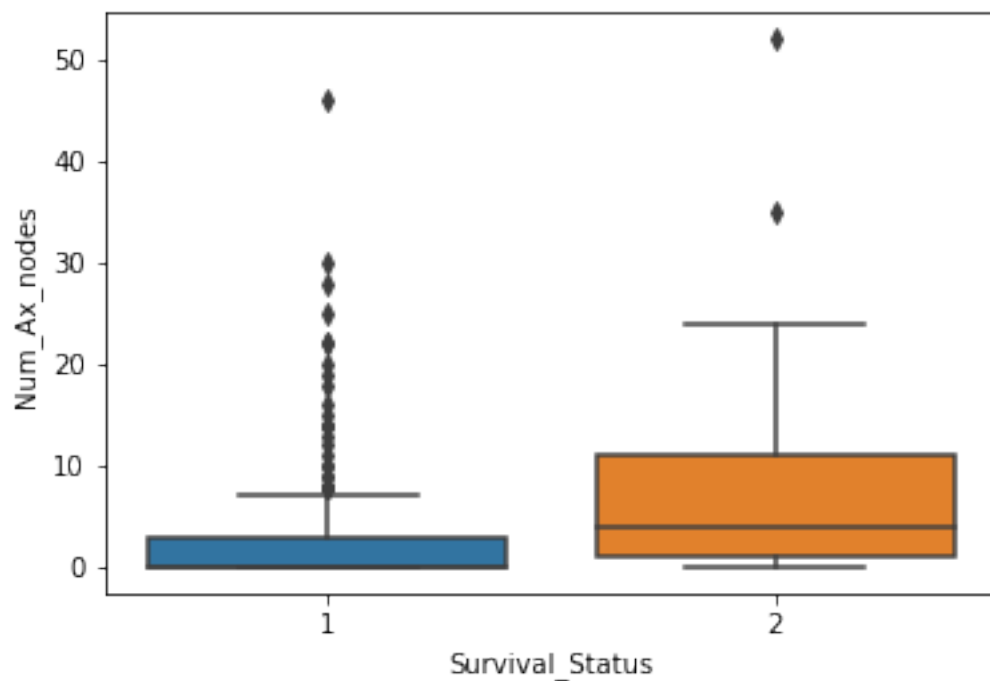
Median Number of Positive Axil Nodes found for class 1 : 30.0

Median Number of Positive Axil Nodes found for class 2 : 34.5

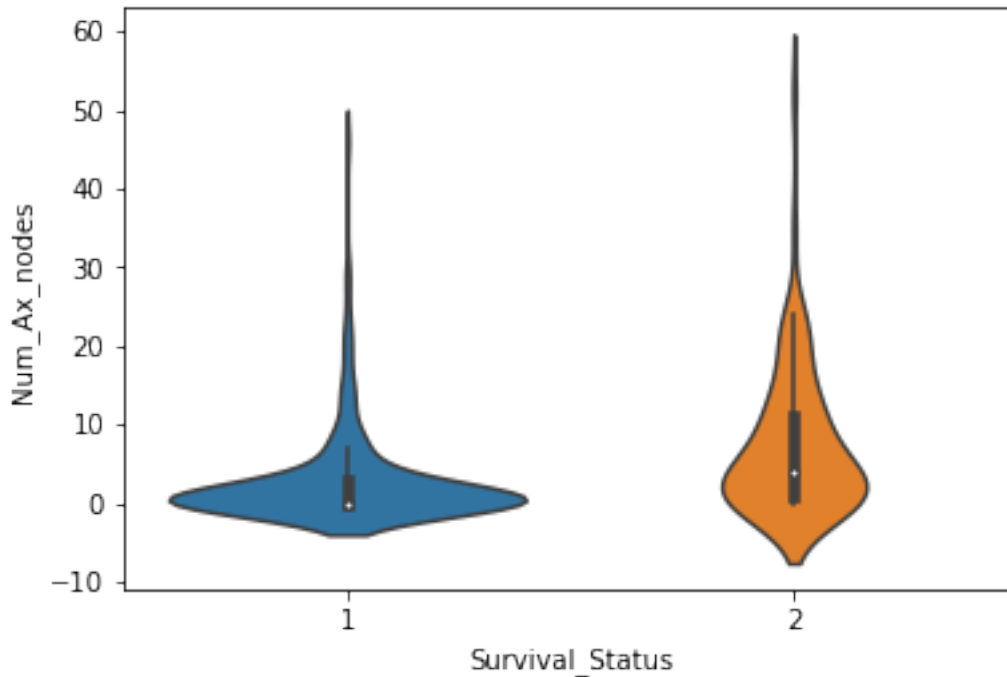
Observations are confusing as the median of both the classes differ only by 4 so it cannot be the spread of the disease that might have been the crucial reason behind their deaths. Next Step trying out Box and Violin plots so understand more about the density of the Number of positive Axial Nodes found

Trying out box plot to see the information visually and to determine the difference between the variance of the Num_Ax_nodes between both classes

```
In [8]: sns.boxplot(y='Num_Ax_nodes', x='Survival_Status', data=haberman)
plt.show()
```



```
In [9]: #Denser regions are darker and sparser regions are thinner
sns.violinplot(y='Num_Ax_nodes', x='Survival_Status', size=10, data=haberman)
plt.show()
```



As is clearly evident from the data is that 75 percentile of the people who could not live more than 5 years had 10 positive axial nodes where are the 75 percentile of the people who could live longer than 5 years is very much lesser (approx <5)

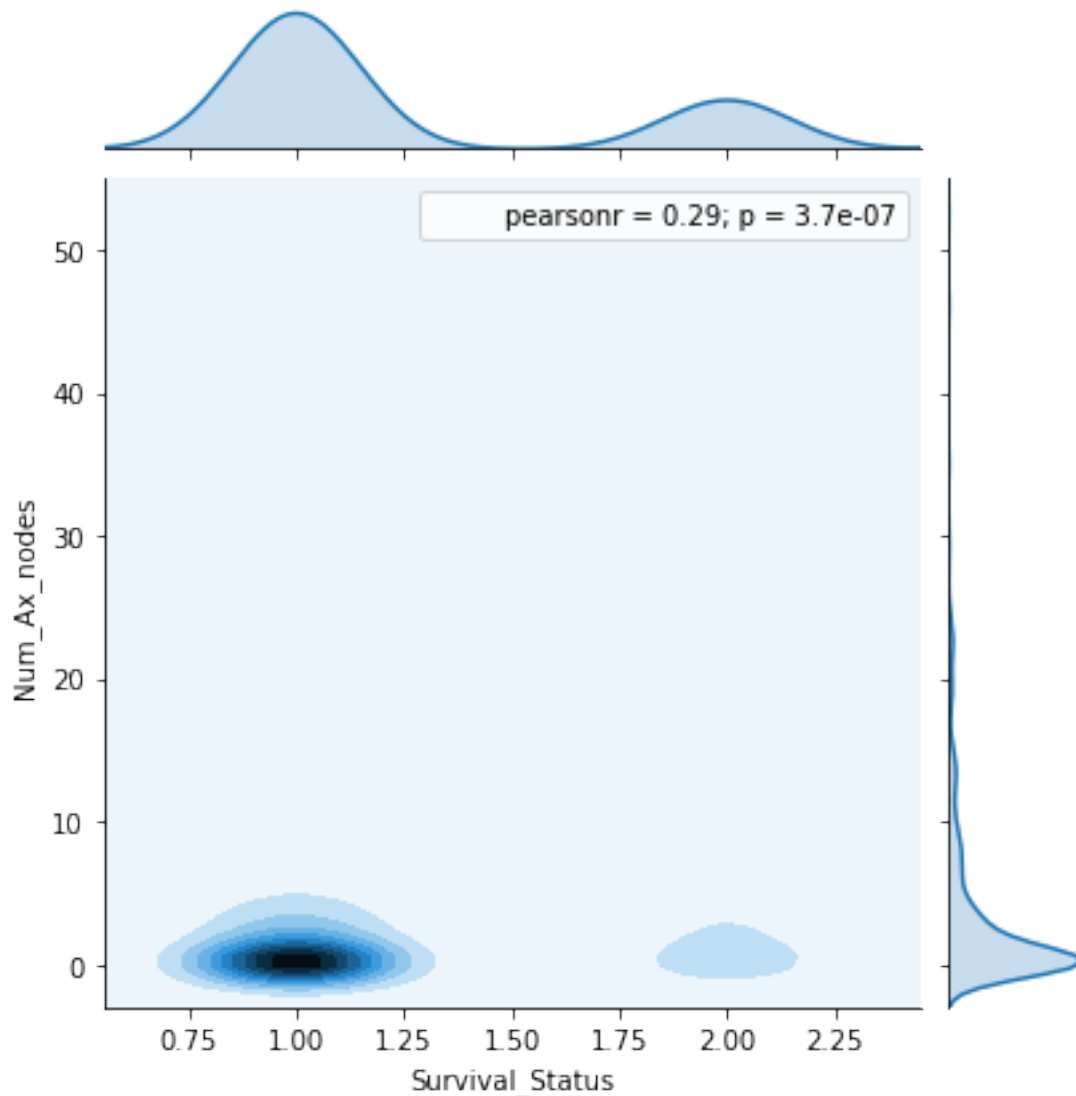
As is also evident from the violin plots is that the median is really very close to the 25 percentile in case 1 and the density of people having lesser number of the positive nodes is higher and is below 10.

But there are more people with more positive nodes in class 2 which can be a reason why most of them could not live for more than 5 years.

Although there are quite some outliers in class 1 who survived to live more than 5 years despite having more number of positive axial nodes

```
In [10]: sns.jointplot(y='Num_Ax_nodes', x='Survival_Status', data=haberman, kind='kde')
```

```
Out[10]: <seaborn.axisgrid.JointGrid at 0x7fe74099fd30>
```



I tried to keep all the experiments that I did but this really confuses me. I cannot interpret this contour plot and its significance. Please respond with remarks and suggestions

0.2 Final Conclusion

The given dataset is an imbalanced dataset.

It is not linearly separable using a simple if-else construct.

The Number of Positive Axill nodes says a lot more about the story as it shows the extent of the spread of the disease so intuitively it is something which helps determining the life expectancy as it is shown above in the data exploratory analysis.