Assignment -1 Haberman Dataset Analysis

(1.1) About Haberman Dataset

- A dataset related to the case study conducted between 1958 and 1970.
- This study was done at University of Chicago's Billings Hospital.
- The study was performed on patients who had undergone breast cancer surgery.
- We find two types of patients in this dataset -

  1) Patient who survived 5 years or longer

  2) Patient who died within 5 years

- Objective: To find whether any of the features present in the dataset like Age,Operation year,Axil nodes are affecting the Survival status of the patient.

In [3]:

```
#importing essential libraries for performing analysis on the dataset
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

import warnings
warnings.filterwarnings("ignore")

haberman = pd.read_csv("haberman.csv")
```

In [4]:

```
#To count the features of the dataset
print (haberman.shape)
```

```
(306, 4)
```

In [5]:

```
#To find the columns present in gicen dataset
print (haberman.columns)
```

```
Index(['Age', 'Op_year', 'axil_nodes', 'Surv_status'], dtype='object')
```

In [6]:

```
# To count the patients that survived or did not survive post operation
haberman["Surv_status"].value_counts()
```

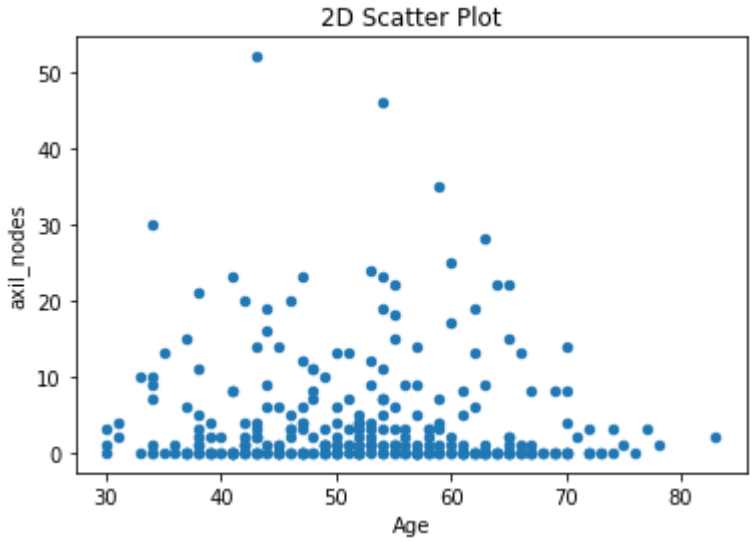
Out[6]:

```
1    225
2     81
Name: Surv_status, dtype: int64
```

(1.2) 2-D Scatter Plot

In [9]:

```
#2-D scatter plot:
#This is to identify whether the age of a person has some relation
#with axillary nodes that are detected in his body

haberman.plot(kind='scatter', x='Age', y ='axil_nodes') ;
plt.title("2D Scatter Plot", fontdict=None, loc='center', pad=None)
plt.show()
```

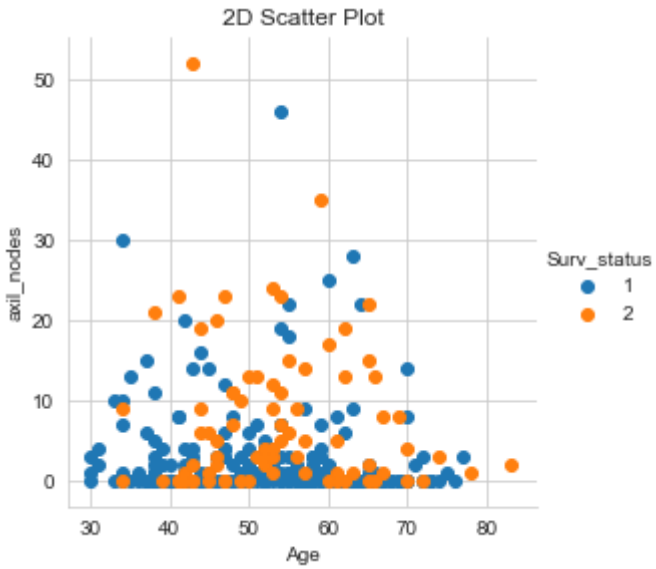


Observations :-

1. More number of axil nodes are foud in patients aged 45 to 65.
2. There are a few outliers lying in the range of 30 to 50.

In [11]:

```
#2-D scatter plot with color coding for each Survival status class
sns.set_style("whitegrid");
sns.FacetGrid(haberman, hue="Surv_status", size=4) \
   .map(plt.scatter, "Age", "axil_nodes") \
   .add_legend();
plt.title("2D Scatter Plot", fontdict=None, loc='center', pad=None)
plt.show();
import warnings
warnings.filterwarnings("ignore")
```



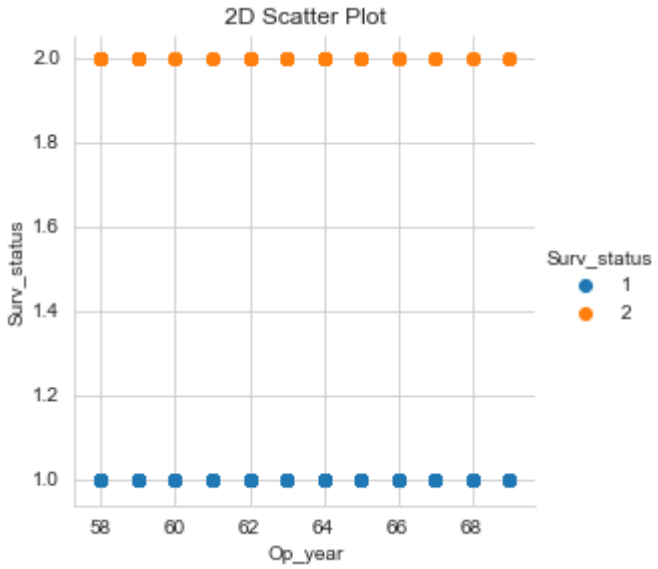
Observations :

1.More patients are found having greater nodes in the age range of 45 to 65.

2.Deaths are higher with persons having nodes 1 to 20.

In [12]:

```
#plotting graph between Survival status & Operation year.
sns.set_style("whitegrid");
sns.FacetGrid(haberman, hue="Surv_status", size=4) \
   .map(plt.scatter, "Op_year", "Surv_status") \
   .add_legend();
plt.title("2D Scatter Plot", fontdict=None, loc='center', pad=None)
plt.show();
```



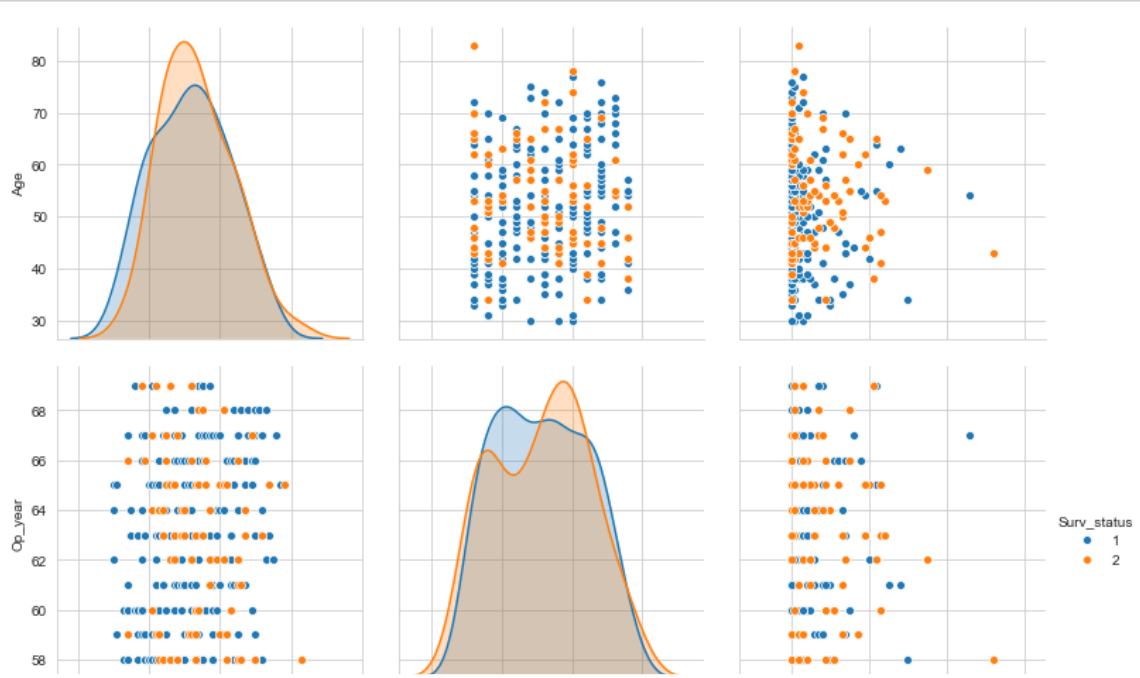
Observations :

It is difficult to find any conclusion from the above graph.

(1.3) Pair-plot

In [14]:

```
# pairwise scatter plot: Pair-plot
plt.close();
sns.set_style("whitegrid");
sns.pairplot(haberman, vars = ['Age', 'Op_year', 'axil_nodes'],
             hue="Surv_status", height=3.5);
plt.show()
```



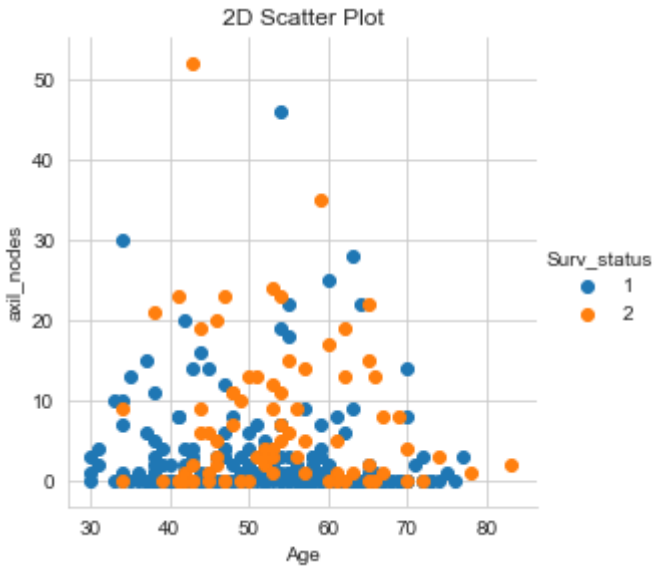
Observations :-

1.The above graphs are very much overlapping with each other so it is quite difficult to come to any strong conclusion based on the same.

2.While the last row of the graphs plotted are very much seperated from ecah other its very difficult to find any relation between the features of the graph.

In [15]:

```
sns.set_style("whitegrid");
sns.FacetGrid(haberman, hue="Surv_status", height=4) \
   .map(plt.scatter, "Age", "axil_nodes") \
   .add_legend();
plt.title("2D Scatter Plot", fontdict=None, loc='center', pad=None)
plt.show();
```



Observations :-

1.The ages which represent greater number of axil nodes are 50-70 with a few cases where it was present for the 30-40yrs aged person.

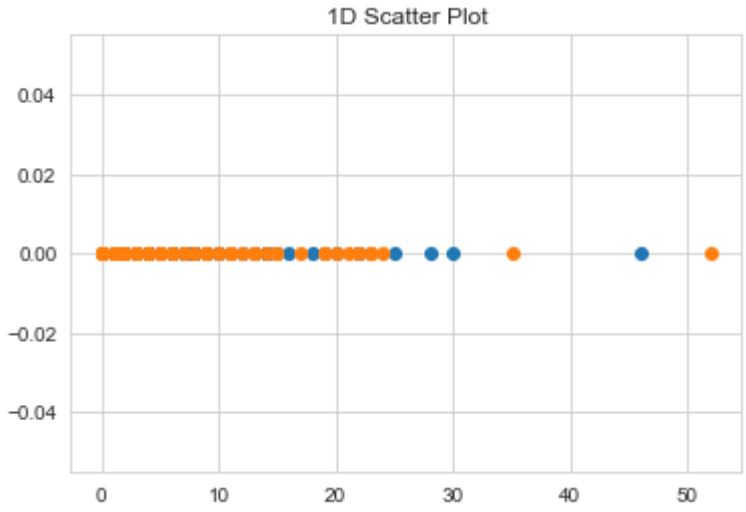
2.Patients with nodes less than 10 had a much higher survival count than the pateints having more than 10 nodes.

3.But also the toll of non surviving patients is higher when the axil nodes are less than 10.

(1.4) Histogram, PDF, CDF

In [16]:

```
#Plotting 1-D scatter plot
#1-D Scatter plot of Survival status
haberman_1 = haberman.loc[haberman["Surv_status"] == 1];
haberman_2 = haberman.loc[haberman["Surv_status"] == 2];
#print()
plt.plot(haberman_1["axil_nodes"], np.zeros_like(haberman_1['axil_nodes']), 'o')
plt.plot(haberman_2["axil_nodes"], np.zeros_like(haberman_2['axil_nodes']), 'o')

plt.title("1D Scatter Plot", fontdict=None, loc='center', pad=None)
plt.show()
```



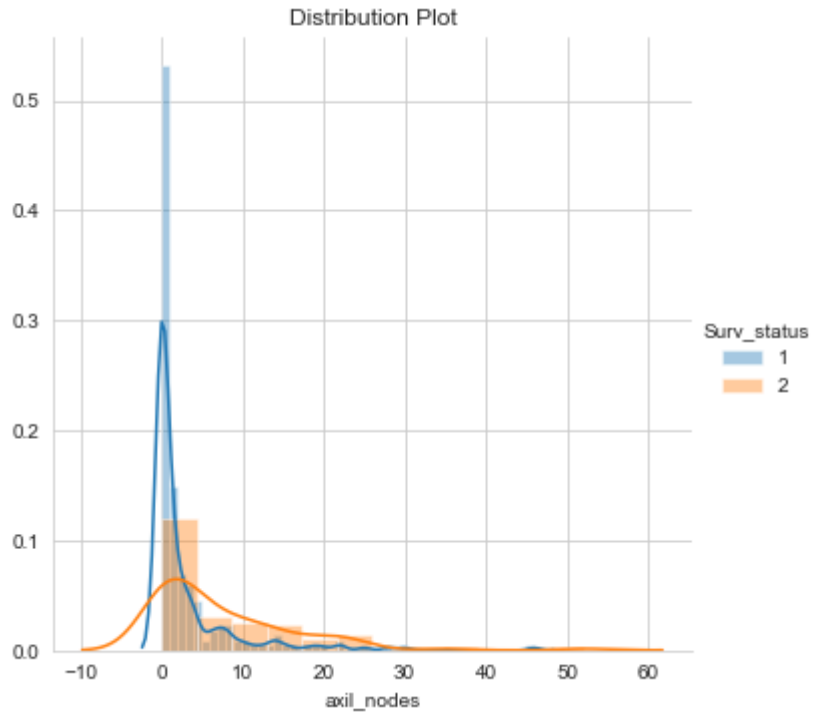
Observations :-

1.As we see very few blue dots in the above graph we can say that much of it would be overlapping between the values of 0 to 20.

2.We can say that both the number of patients surviving & not surviving are higher with axil nodes less than 25.

In [17]:

```
#plotting for comparison between survival status & axil nodes
sns.FacetGrid(haberman, hue="Surv_status", size=5) \
   .map(sns.distplot, "axil_nodes") \
   .add_legend();
plt.title("Distribution Plot", fontdict=None, loc='center', pad=None)
plt.show();
```



Observations :-

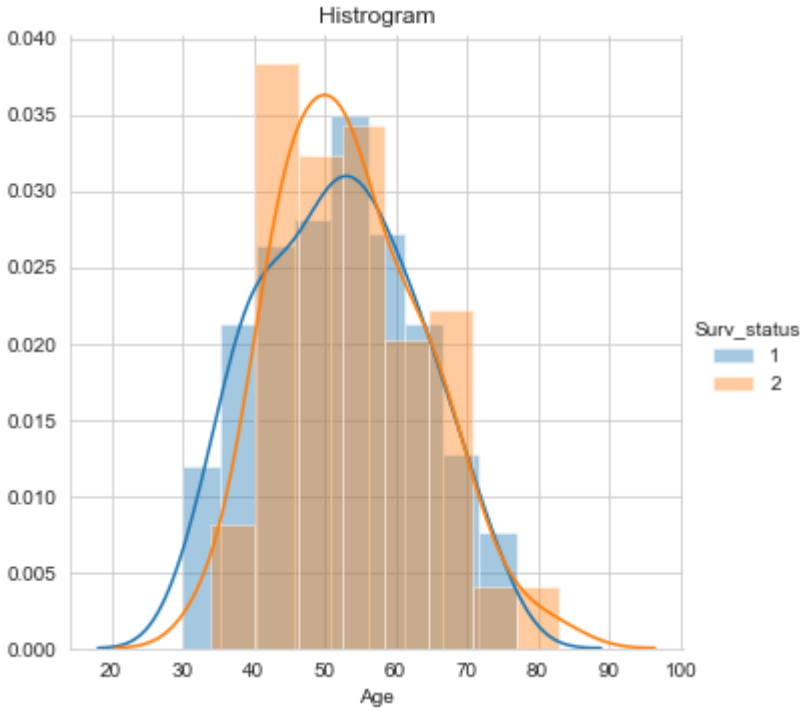
1.As previously mentioned this graph shows similar observation of number of survival & deaths both being higher for axil nodes of 0 to 10.

2.The graphs for Survival & deaths both being very much overlapping it is very difficult to find any conclusion.

3.The Y-axis of the graph seems to be irrelevant in the above graph.

In [19]:

```
#Plotting for comparsion of survival status & Age
sns.FacetGrid(haberman, hue="Surv_status", size=5) \
   .map(sns.distplot, "Age") \
   .add_legend();
plt.title("Histrogram", fontdict=None, loc='center', pad=None)
plt.show();
```



Observations :-

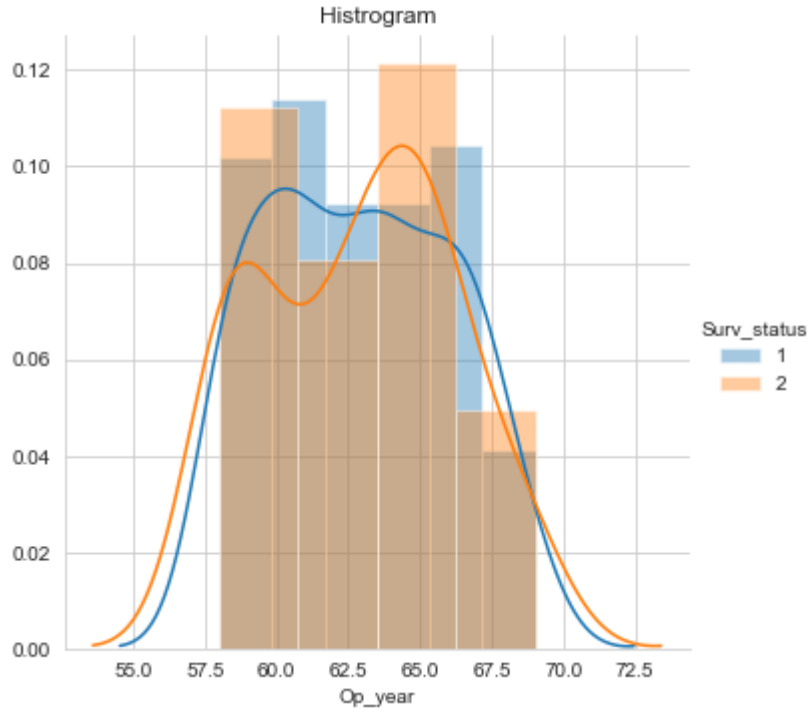
1.The number of Survival case & death case both higher for the age of 40 to 70.

2.The death toll seems to be more in if considered the age group of 40 to 50 whereas the survival rate is slightly higher or similar for age group of 50 to 65.

3.Survival status is higher for age group of 70-75.

In [20]:

```
#plotting histogram for operation year
sns.FacetGrid(haberman, hue="Surv_status", size=5) \
   .map(sns.distplot, "Op_year") \
   .add_legend();
plt.title("Histrogram", fontdict=None, loc='center', pad=None)
plt.show();
```



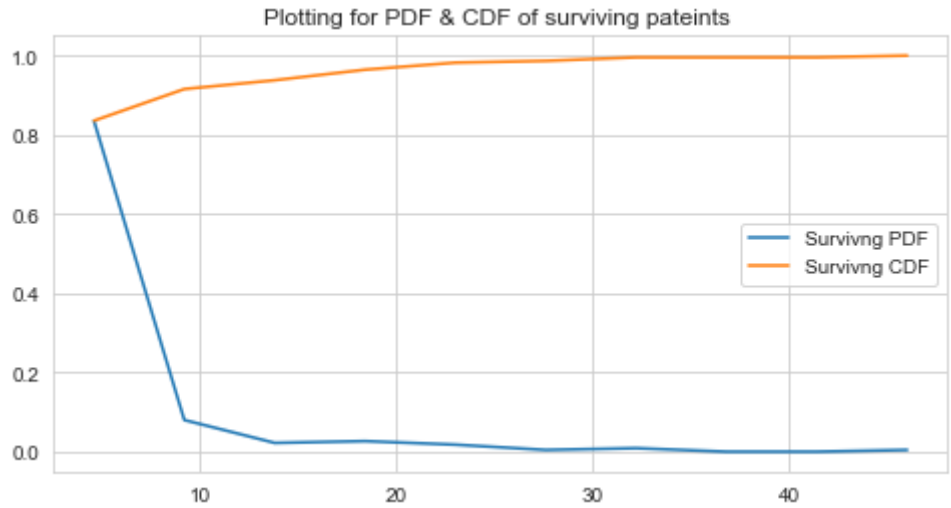
Observation :-

1.The above graph is plotted between Surv status & operation year but it is very difficult to find any meaningful insight from the above graph.

In [23]:

```
#plotting CDF & PDF for axial nodes of surviving patients

fig, ax = plt.subplots(figsize=(8, 4))
counts, bin_edges = np.histogram(haberman_1['axil_nodes'], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges)

#compute cdf
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.legend()

plt.legend(['Not_sur_pdf','Not_sur_cdf','Sur_pdf','Sur_cdf'])
plt.legend(['Survivng PDF','Survivng CDF'])
plt.title("Plotting for PDF & CDF of surviving pateints ", fontdict=None, loc='center', pad
plt.show();

import warnings
warnings.filterwarnings("ignore")
```

```
No handles with labels found to put in legend.

[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. ]
```



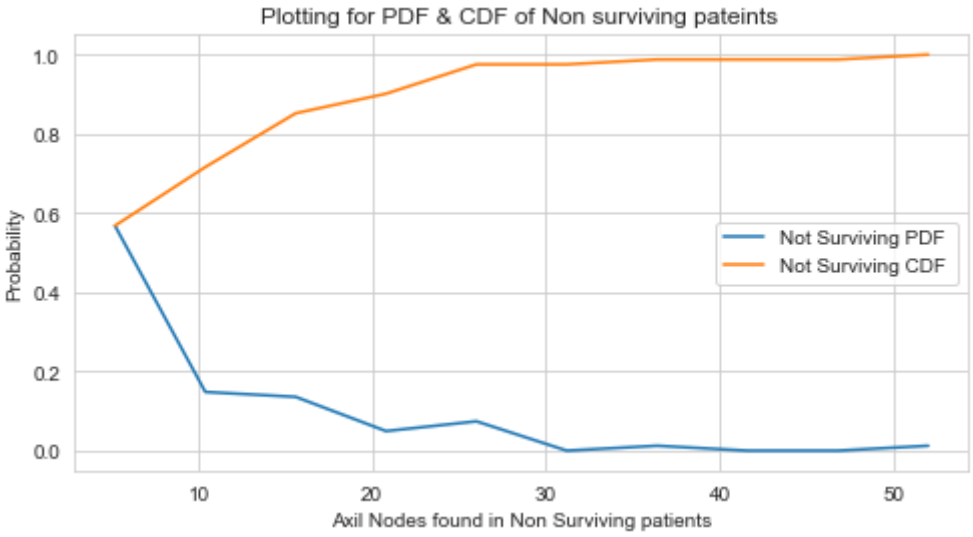
Observations :-

1. The above cdf & pdf conclude that 82% of the patients have axil nodes less than 10.
2. 95% of the survived patients have axil nodes less than or equal to 20.

In [24]:

```
#plotting CDF & PDF for axial nodes of Nonsurviving patients
fig, ax = plt.subplots(figsize=(8, 4))
counts, bin_edges = np.histogram(haberman_2['axil_nodes'], bins=10,
                                 density = True)

pdf = counts/(sum(counts))
print(pdf);
print(bin_edges)

#compute cdf
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)

ax.set_ylabel('Probability')
ax.set_xlabel('Axil Nodes found in Non Surviving patients')
plt.legend(['Not_sur_pdf','Not_sur_cdf','Sur_pdf','Sur_cdf'])
plt.legend(['Not Surviving PDF ','Not Surviving CDF'])
plt.title("Plotting for PDF & CDF of Non surviving pateints ", fontdict=None, loc='center',
plt.show();
```

```
[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. ]
```



Observations :-

1. 58% of the patients who died had axil nodes equal to 4.
2. 87% of the patients who died had axil nodes less than or equal to 20.

In [16]:

```
#Mean, Variance, std-deviation
print("Means:")
print(np.mean(haberman_1["axil_nodes"]))
#Mean with an outlier
print(np.mean(np.append(haberman_1["axil_nodes"],50)));
print(np.mean(haberman_2["axil_nodes"]))

print("\nStd-dev:");
print(np.std(haberman_1["axil_nodes"]))
print(np.std(haberman_2["axil_nodes"]))
```

```
Means:
2.7911111111111113
3.0
7.45679012345679

Std-dev:
5.857258449412131
9.128776076761632
```

In [17]:

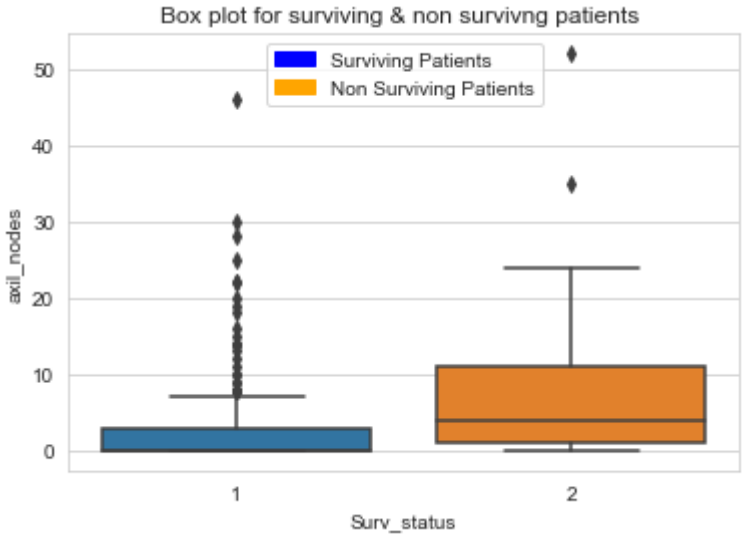
```
#Median, Qunatiles, Percentiles, IQR
print("\nMedians:")
print(np.median(haberman_1["axil_nodes"]))
#Median with an outlier
print(np.median(np.append(haberman_1["axil_nodes"],50)));
print(np.median(haberman_2["axil_nodes"]))

print("\nQuantiles")
print(np.percentile(haberman_1["axil_nodes"],np.arange(0, 100, 25)))
print(np.percentile(haberman_2["axil_nodes"],np.arange(0,100,25)))

print("\n90th Percentiles:")
print(np.percentile(haberman_1["axil_nodes"],90))
print(np.percentile(haberman_2["axil_nodes"],90))

from statsmodels import robust
print("\nMedian Absolute Deviation")
print(robust.mad(haberman_1["axil_nodes"]))
print(robust.mad(haberman_2["axil_nodes"]))
```

```
Medians:
0.0
0.0
4.0

Quantiles
[0. 0. 0. 3.]
[ 0.  1.  4. 11.]

90th Percentiles:
8.0
20.0

Median Absolute Deviation
0.0
5.930408874022408
```

Box plot & whiskers

In [26]:

```
#Plotting box plot for surviving & Non surviving patients
import matplotlib.patches as mpatches
colour=['blue','orange']
fig, ax = plt.subplots()

sns.boxplot(x='Surv_status',y='axil_nodes', data=haberman)

blue_patch = mpatches.Patch(color='blue', label='Surviving Patients')
orange_patch = mpatches.Patch(color='orange', label='Non Surviving Patients')
plt.legend(handles=[blue_patch, orange_patch])
plt.title("Box plot for surviving & non survivng patients ", fontdict=None, loc='center', p

plt.show()
```

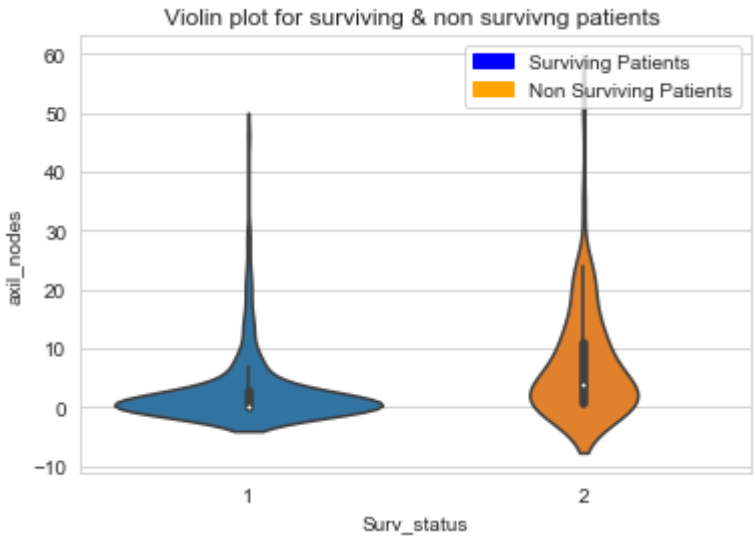


Observations :

1. 25th percentile value & 50th percentile value for the survival cases have the same value i.e. 0.
2. 75th percentile value of the survival cases lies between 25th & 50th percentile of the non surviving patients.
3. 75th percentile value of surviving case is equal to 3.
4. Many outliers can be seen in the surviving case which lies in the range of 8 to 31.
5. While for the non surviving patients there are very few outliers.

In [27]:

```
#Plotting violin plot for surviving & Non surviving patients
colour=['blue','orange']
fig, ax = plt.subplots()

sns.violinplot(x="Surv_status", y="axil_nodes", data=haberman, size=8)


blue_patch = mpatches.Patch(color='blue', label='Surviving Patients')
orange_patch = mpatches.Patch(color='orange', label='Non Surviving Patients')
plt.legend(handles=[blue_patch, orange_patch])
plt.title("Violin plot for surviving & non survivng patients ", fontdict=None, loc='center'
plt.show()
```



Observations :

1. The Box plot box range in case 1 seems to be very confined to come to an conclusion.
2. The Box plot box range in case 2 varies from 0 to 12.
3. The whiskers in both the violin plots is quite wide as we have seen similarly in box plot previously plotted.

Key Findings / Conclusions :-

1. Clearly from all the above analysis are the featues of axil nodes & Surv status are the heart of this dataset.
2. The axil node is the only factor that could possibly have relation with the survival & death of the person.
3. But in the above analysis we come to know that the patients having less number of axil nodes have survived & even not survived in many cases.So the plots with these featues are very much overlapping with each other to find any strong conclusion.
4. The age factor of a patient also gives mixed results as aove to find any strong connection between survival status of a patient.
5. The rest of the features in the dataset seem to be irrelevant with the survival case of the patients.

In [ ]: