Haberman Cancer Survival Data-set Using EDA

Objective:- To predict the patient will survive after 5 years or not on the basis of patient's age, operation treatment year and the number of axillary nodes.

Explanation about features:-

- Age:- Age of the patient at the time of operation.
- Operation_Year:- Patients year of operation.
- axil_nodes:- Number of positive axillary nodes.
- Surv_status:-Survival Status, 1 = the patient survived 5 years or longer, 2 = the patient died within 5 year

In [5]:

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

In [13]:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

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

In [6]:

```
#Data points
print(haberman.shape)
```

```
(306, 4)
```

In [10]:

```
#Number of features
print(haberman.columns)
```

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

In [8]:

```
#Number of classes
#Data points per class
haberman['Surv_status'].value_counts()
```

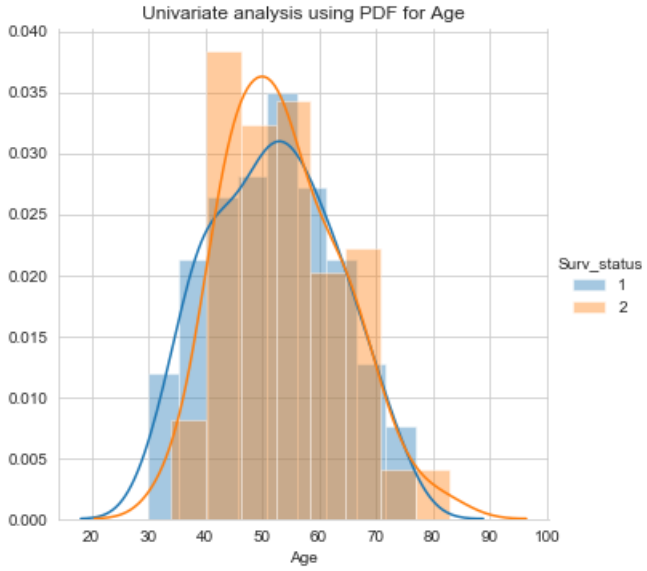
Out[8]:

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

Univariate Analysis using PDF

In [30]:

```
sns.FacetGrid(haberman, hue="Surv_status", size=5)\
    .map(sns.distplot, "Age")\
    .add_legend();

plt.title("Univariate analysis using PDF for Age")
plt.show();
```

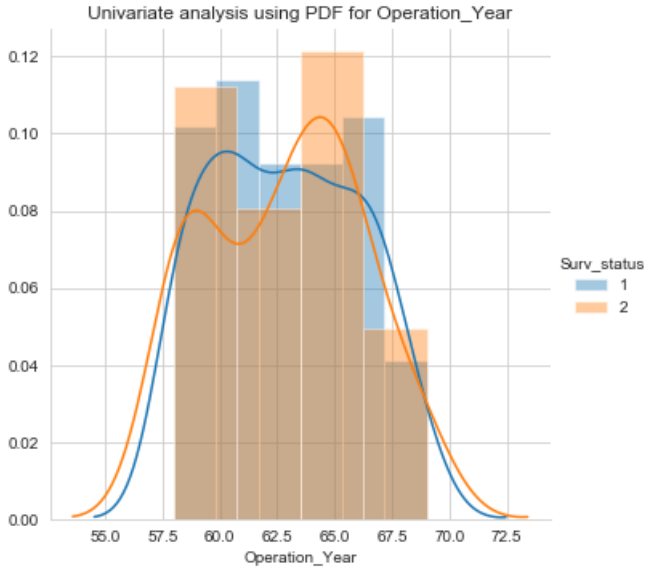


Observation :-

1. Most of the patients survived in 5 years are in the age of range between 30 to 75
2. The patients died within 5 years are in the age of range between 40 to 70

In [31]:

```
sns.FacetGrid(haberman, hue="Surv_status", size=5)\
    .map(sns.distplot, "Operation_Year")\
    .add_legend();
plt.title("Univariate analysis using PDF for Operation_Year")
plt.show();
```

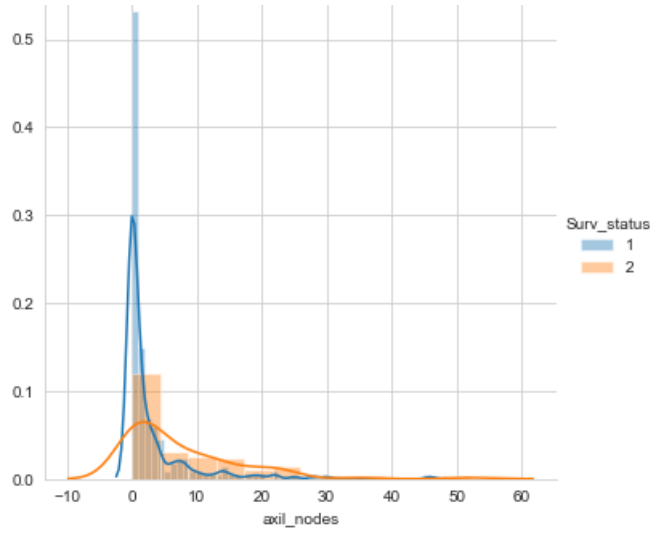


Observation :-

1. Most of the patients survived in 5 years are in the Operation Year of range between 58 to 68
2. The patients died within 5 years are in the Operation Year of range between 58 to 68

In [32]:

```
sns.FacetGrid(haberman, hue="Surv_status", size=5)\
    .map(sns.distplot, "axil_nodes")\
    .add_legend();
plt.title("Univariate analysis using PDF for axil_nodes")
plt.show();
```



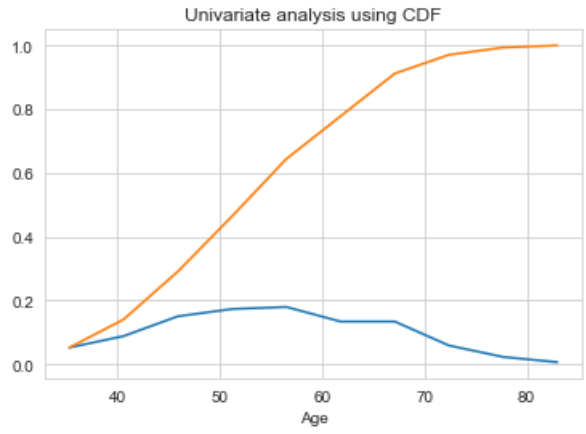
Observation:- Axillary nodes can be easily classified from other two features.

Univariate analysis using CDF

In [45]:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

haberman = pd.read_csv('haberman.csv')

counts, bin_edges = np.histogram(haberman['Age'], 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.xlabel("Age");
plt.title("Univariate analysis using CDF")
plt.show();
```

```
[0.05228758 0.08823529 0.1503268  0.17320261 0.17973856 0.13398693
 0.13398693 0.05882353 0.02287582 0.00653595]
[30.  35.3 40.6 45.9 51.2 56.5 61.8 67.1 72.4 77.7 83. ]
```



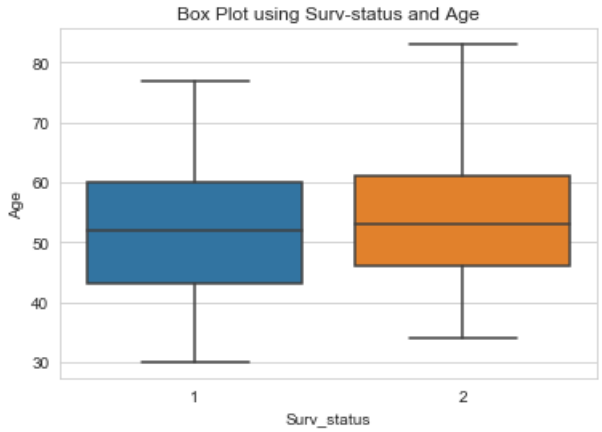
Observation:-

- PDF represents 20% points between the age of 50 to 80
- CDF represents 70% points less than Age 60 (i.e Age<=60) as well as 100% points indicates that less than Age 80

Box Plot

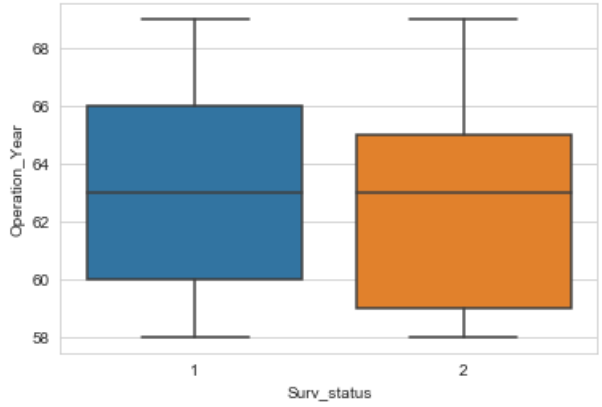
In [28]:

```
sns.boxplot(x='Surv_status', y='Age', data=haberman)
plt.title("Box Plot using Surv-status and Age")
plt.show()
```



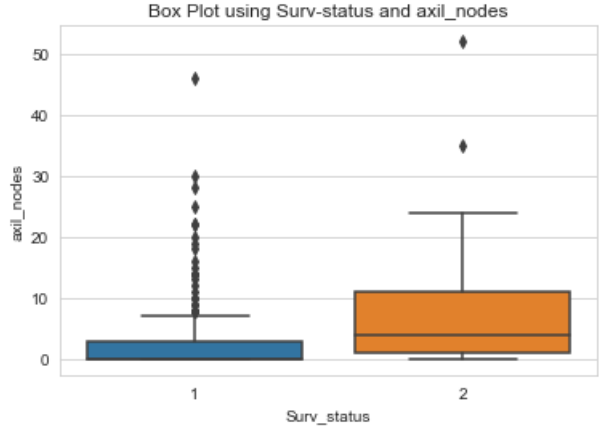
In [51]:

```
sns.boxplot(x='Surv_status', y='Operation_Year', data=haberman)
plt.title("Box Plot using Surv-status and Operation_year")
plt.show()
```



In [34]:

```
sns.boxplot(x='Surv_status', y='axil_nodes', data=haberman)
plt.title("Box Plot using Surv-status and axil_nodes")
plt.show()
```



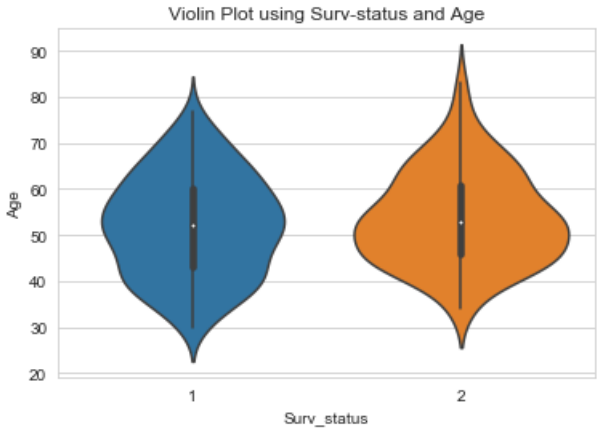
Observation:-

1. According to axillary nodes of Surv_status 1, the patients survived are in the range 0 t 3
2. According to axillary nodes of Surv_status 2, the patient died are in the range 2 to 10

Violin Plot

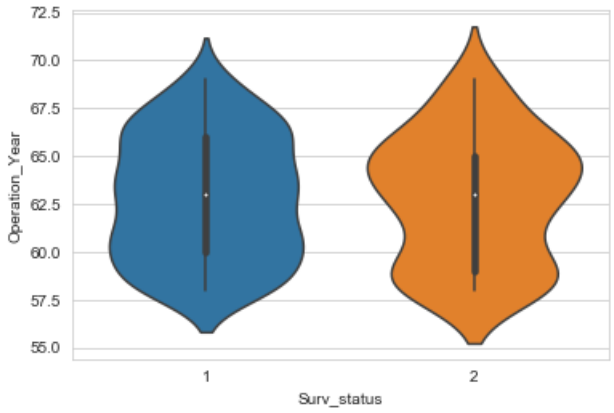
In [36]:

```
sns.violinplot(x='Surv_status', y='Age', data=haberman)
plt.title("Violin Plot using Surv-status and Age")
plt.show()
```



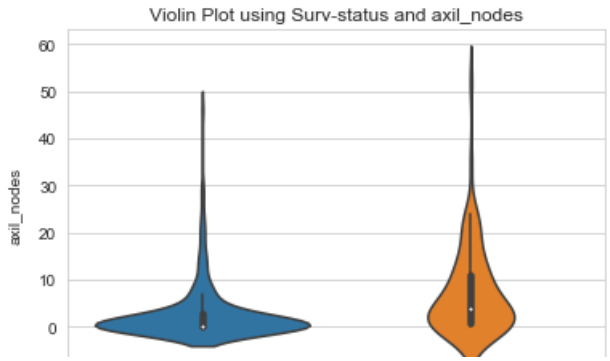
In [53]:

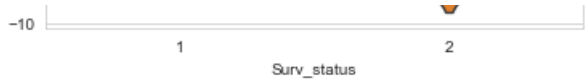
```
sns.violinplot(x='Surv_status', y='Operation_Year', data=haberman)
plt.title("Violin Plot using Surv-status and Operation_year")
plt.show()
```



In [37]:

```
sns.violinplot(x='Surv_status', y='axil_nodes', data=haberman)
plt.title("Violin Plot using Surv-status and axil_nodes")
plt.show()
```





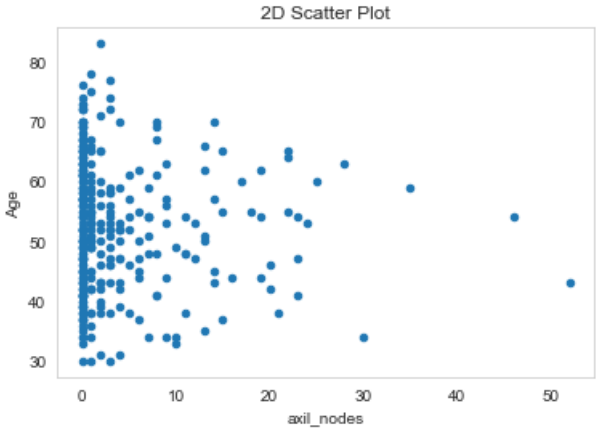
Observation:-

1. The number of axillary nodes of the survived patients is from 0 to 5.

2D Scatter Plot

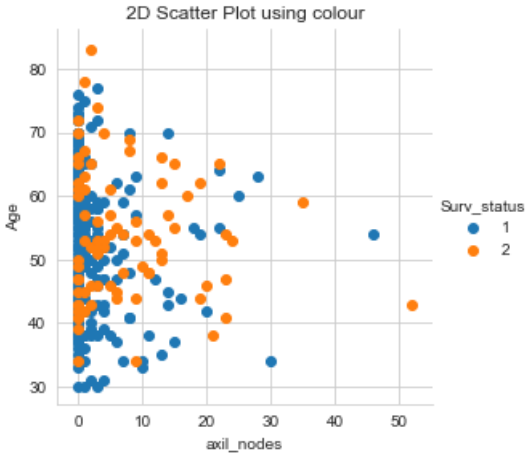
In [40]:

```
haberman.plot(kind="scatter",x="axil_nodes",y="Age");
plt.grid()
plt.title("2D Scatter Plot")
plt.show()
```



In [41]:

```
#2d scatter plot using colour

sns.set_style("whitegrid");
sns.FacetGrid(haberman, hue="Surv_status", size=4) \
    .map(plt.scatter,"axil_nodes", "Age")\
    .add_legend();
plt.title("2D Scatter Plot using colour")
plt.show();
```



Observation :-

1. According to 2D scatter plot, patients survived in 5 years are in the age of range between 30 to 75
2. The patients died within 5 years are in the age of range between 40 to 70

Pair Plot

In [26]:

```
sns.pairplot(haberman, vars=["Age", "Operation_Year", "axil_nodes"], size=3,\
    hue="Surv_status")
plt.show()
```

Observation:-

1. Operation_Year and axillary nodes are the very useful features to identify patients survival from cancer.
2. The patients survived in 5 years are in the operation year of range between 60 to 70
3. The patients died in 5 years are in the operation year of range between 55 to 70

Conclusion:-

1. 75% of patients are survived in 5 years
2. 25% of patients died within 5 years
3. Haberman Datasets are imbalanced dataset.