In [2]:

```
#Objective which factor affects the most on the survival of patient undergone surgery for breast c
ancer

#Codes are taken from ipython note book provided under video lecture

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

haberman = pd.read_csv("C:\Applied AI\Exploratory Data Analysis\haberman.csv")

print (haberman.shape)
```

```
(306, 4)
```

In [3]:

```
print (haberman.columns)
```

```
Index(['Age', 'Operation Year', 'Axil nodes', 'Status'], dtype='object')
```

In [4]:

```
haberman["Status"].value_counts()
```

Out[4]:

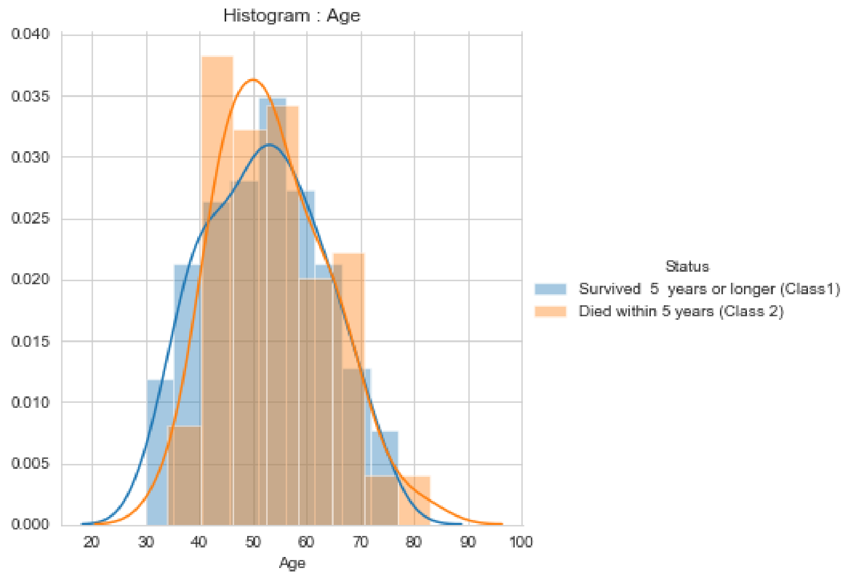
```
Survived  5  years or longer (Class1)    225
Died within 5 years (Class 2)             81
Name: Status, dtype: int64
```

Observation:

The data set is imbalanced as the patient labelled as class 1 and Class 2 based on their survival differs hugely

In [25]:

```
sns.set_style("whitegrid");
sns.FacetGrid(haberman, hue="Status", height=5) \
   .map(sns.distplot, "Age") \
   .add_legend();
plt.title('Histogram : Age')
plt.show();
```



Observations:

1.Between 30 to 40 more patients have survived beyond 5 years

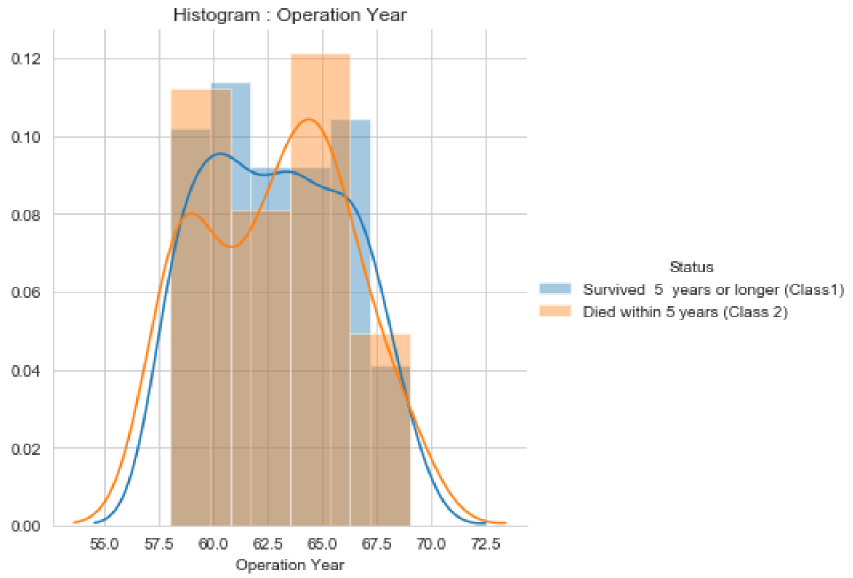
2.Between 40 to 60 more patients have died within 5 years

3.Between 60 to 75 the chances are equiprobable.

4.Above 75 the chance of dieying within 5 years is higher.

In [26]:

```
sns.FacetGrid(haberman, hue="Status", height=5) \
   .map(sns.distplot, "Operation Year") \
   .add_legend();
plt.title('Histogram : Operation Year')
plt.show()
```



Observation:

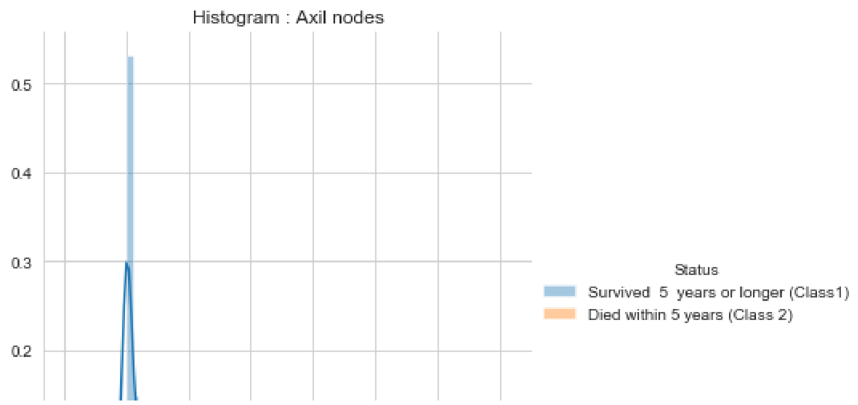
1.Between 1958 to 1962 the chance of surviving beyond 5 years is high.

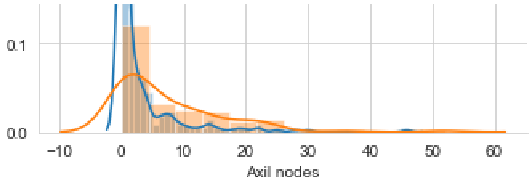
2.Between 1962 to 1965 chance of dieying with 5 years is high.

3.Beyond 1965 there is no significant difference in the survival between two clases.

In [28]:

```
sns.FacetGrid(haberman, hue="Status", height=5) \
   .map(sns.distplot, "Axil nodes") \
   .add_legend();
plt.title('Histogram : Axil nodes')
plt.show()
```





Observation

1.When the affected nodes is 0 the chances of surviving beyond 5 years is higher.

2.As the number of nodes increases the chances of dying within 5 years increases.

Pair Plot

In [13]:

```
sns.pairplot(haberman, hue="Status", height=3)
plt.show()
```

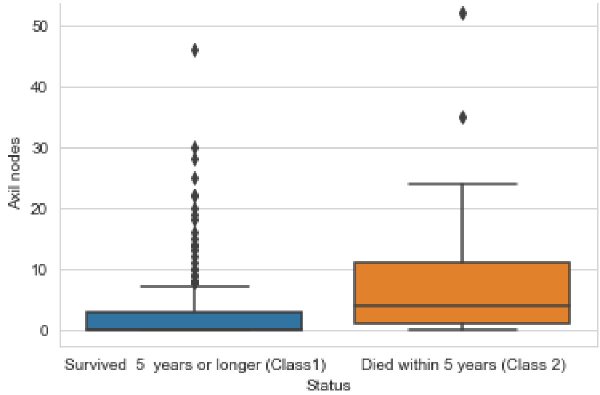
Observation:

Axil nodes provide better understanding on patients survival

In [14]:

```
sns.boxplot(x='Status',y='Axil nodes', data=haberman)
plt.title('Box Plot and Whiskers')
plt.show()
```

Box Plot and Whiskers



Observation:

1.As the number of Axil nodes increases the chances of dieying with in 5 years increases

2.50% of patients who died with in 5 years seems to have atleast 4 Axil node.The result of 4 Axil nodes approximately matches with the mean Axil nodes calculated below.

3.Where as 50% of patients who survided 5 years or longer got 0 Axil nodes

In [16]:

```
sns.violinplot(x="Status", y="Axil nodes", data=haberman, size=5)
plt.title('Violin Plot')
plt.show()
```

Observation:

When the number of Axil nodes affected is less the chance of surviving beyond 5 years is higher.

In [17]:

```
print("Age:Mean & Standard Deviation")
print(np.mean(haberman["Age"]))
print(np.std(haberman["Age"]))
```

```
Age:Mean & Standard Deviation
52.45751633986928
10.78578520363183
```

Observation:

The average age of people undergoing for opration is 52

In [18]:

```
print("Axil nodes:Mean & Standard Deviation")
print(np.mean(haberman["Axil nodes"]))
print(np.std(haberman["Axil nodes"]))
```

```
Axil nodes:Mean & Standard Deviation
4.026143790849673
7.177896092811152
```

Observation:

On an average 4 Axil nodes of the patients were affected.Axil nodes determines the spread of cancer in patients body.

In [19]:

```
print("Age:Medians,Percentile & Quantile")
print(np.median(haberman["Age"]))
print(np.percentile(haberman["Age"],95))
print(np.percentile(haberman["Age"],np.arange(0, 100, 25)))
```

```
Age:Medians,Percentile & Quantile
52.0
70.0
[30.   44.   52.   60.75]
```

Observation:

1. The median age of patients is 52
2. 95% of patient's age is less than 70
3. 25%, 50% , 75% of patient's age are less than 44 , 52 , 60 respectively.

In [20]:

```
print("Axil nodes:Medians,Percentile & Quantile")
print(np.median(haberman["Axil nodes"]))
print(np.percentile(haberman["Axil nodes"],np.arange(0, 100, 5)))
print(np.percentile(haberman["Axil nodes"],np.arange(0, 100, 25)))
```

```
Axil nodes:Medians,Percentile & Quantile
1.0
[ 0.    0.    0.    0.    0.    0.    0.    0.    0.    1.    1.    1.
  2.    3.    3.    4.    7.    9.   13.   19.75]
[0. 0. 1. 4.]
```

Observation:

1. On an average the number of nodes affected in a patient is 1.
2. 40 % of patient's nodes were unaffected.
3. At least 1 node is affected for 60 % of patients.

In [21]:

```
from statsmodels import robust
print ("Median Absolute Deviation : Age & Axil nodes")
print(robust.mad(haberman["Age"]))
print(robust.mad(haberman["Axil nodes"]))
```

```
Median Absolute Deviation : Age & Axil nodes
11.860817748044816
1.482602218505602
```

In [24]:

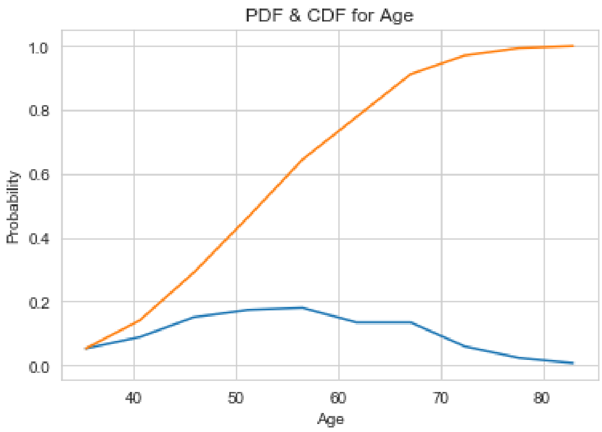
```
counts, bin_edges = np.histogram(haberman ['Age'], bins=10, density = True)

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

```
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)
plt.title('PDF & CDF for Age')
plt.xlabel("Age")
plt.ylabel("Probability")
```

```
[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. ]
```

Out[24]:

```
Text(0, 0.5, 'Probability')
```



Observation:

1. More number patients are in the age group between 50 to 60.
2. 75% of pateient's age is below 60.

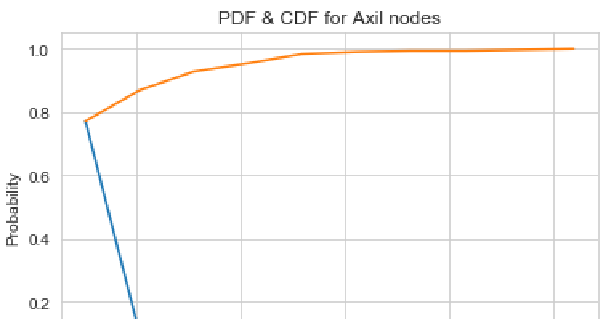
In [23]:

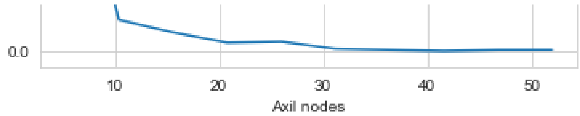
```
counts, bin_edges = np.histogram(haberman ['Axil nodes'], bins=10,density = True)

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

cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.title('PDF & CDF for Axil nodes')
plt.xlabel("Axil nodes")
plt.ylabel("Probability")
```

```
[0.77124183 0.09803922 0.05882353 0.02614379 0.02941176 0.00653595
 0.00326797 0.         0.00326797 0.00326797]
[ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
```

Out[23]:

```
Text(0, 0.5, 'Probability')
```





Observation:

1. Around 2 Axil nodes were affected in about 78% of patients.
2. More than 2 Axil nodes were affected for remaining 22% of patient.