3. Plotting for Exploratory data analysis (EDA)

(3.1) Basic Terminology

- What is EDA?
- Data-point/vector/Observation
- Data-set.
- Feature/Variable/Input-variable/Dependent-varibale
- Label/Indepdendent-variable/Output-varible/Class/Class-label/Response label
- Vector: 2-D, 3-D, 4-D,.... n-D

Q. What is a 1-D vector: Scalar

haberman

In [30]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [31]:
```
haberman = pd.read_csv("haberman.csv")
```

In [32]:
```
print (haberman.shape)
```

```
(305, 4)
```

In [33]:

```
haberman.columns = ["Age","Operation_Year","positive_lymph_nodes","Survival status"
]
print (haberman.columns)
haberman
```

```
Index(['Age', 'Operation_Year', 'positive_lymph_nodes', 'Survival status'], dtyp
e='object')
```

Out[33]:

	Age	Operation_Year	positive_lymph_nodes	Survival status
0	30	62	3	1
1	30	65	0	1
2	31	59	2	1
3	31	65	4	1
4	33	58	10	1
5	33	60	0	1
6	34	59	0	2
7	34	66	9	2
8	34	58	30	1
9	34	60	1	1
10	34	61	10	1
11	34	67	7	1
12	34	60	0	1
13	35	64	13	1
14	35	63	0	1
15	36	60	1	1
16	36	69	0	1
17	37	60	0	1
18	37	63	0	1
19	37	58	0	1
20	37	59	6	1
21	37	60	15	1
22	37	63	0	1
23	38	69	21	2
24	38	59	2	1
25	38	60	0	1
26	38	60	0	1
27	38	62	3	1
28	38	64	1	1
29	38	66	0	1
...	...	...	...	...
275	67	66	0	1
276	67	61	0	1
277	67	65	0	1
278	68	67	0	1
279	68	68	0	1
280	69	67	8	2
281	69	60	0	1
282	69	65	0	1
283	69	66	0	1

In [34]:

```
haberman.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 305 entries, 0 to 304
Data columns (total 4 columns):
Age                      305 non-null int64
Operation_Year           305 non-null int64
positive_lymph_nodes     305 non-null int64
Survival status          305 non-null int64
dtypes: int64(4)
memory usage: 9.6 KB
```

In [35]:

```
haberman.describe()
```

Out[35]:

	Age	Operation_Year	positive_lymph_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

objective

To predict the survival of the patient based on the age,his/her operation_year and the number of positive lymph nodes.

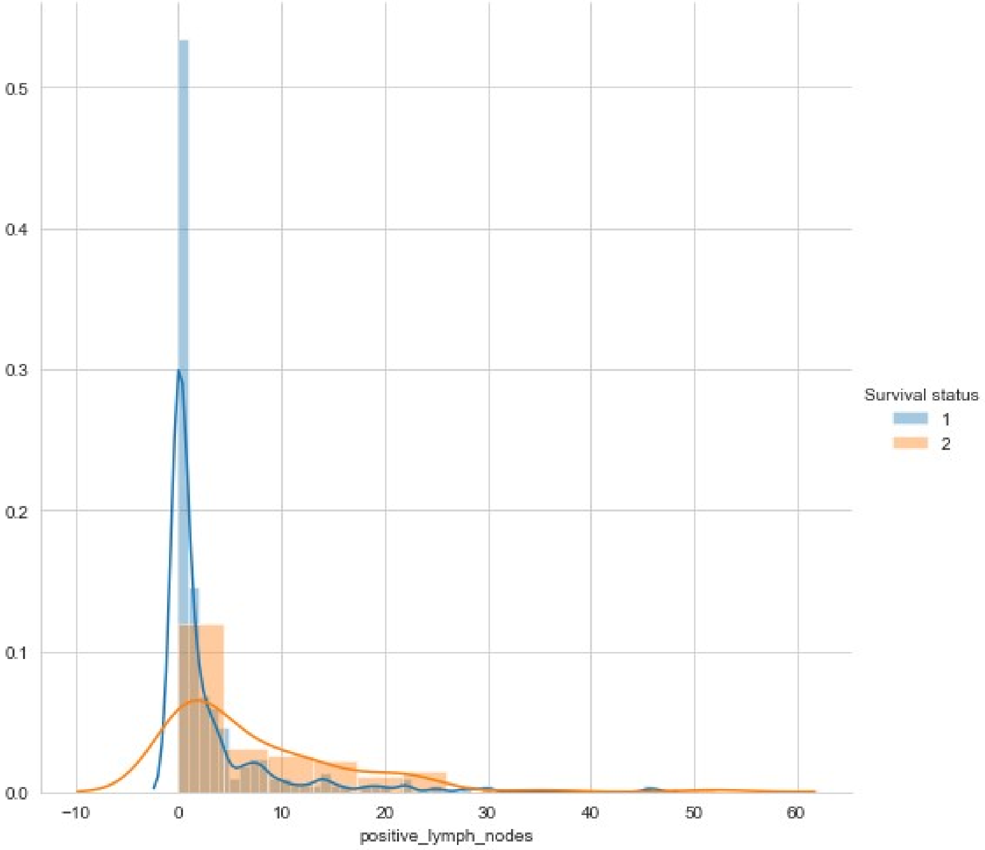
UNIVARIATE ANALYSIS

Histogam, PDF

In [36]:

```
sns.FacetGrid(haberman, hue="Survival status", size=7) \
   .map(sns.distplot, "positive_lymph_nodes") \
   .add_legend();
plt.show();
```

```
C:\Users\Sai charan\Anaconda3\lib\site-packages\seaborn\axisgrid.py:230: UserWar
ning: The `size` paramter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
C:\Users\Sai charan\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: Futur
eWarning: Using a non-tuple sequence for multidimensional indexing is deprecated
; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interp
reted as an array index, `arr[np.array(seq)]`, which will result either in an er
ror or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

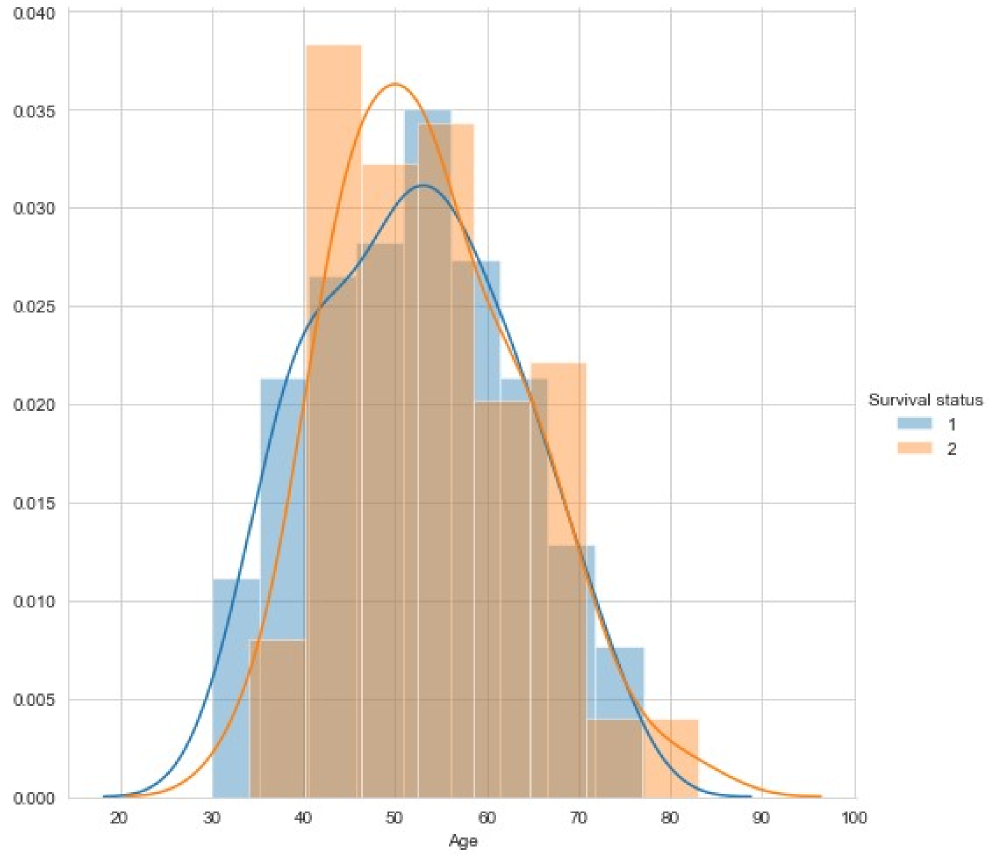


In [ ]:

In [37]:

```
sns.FacetGrid(haberman, hue="Survival status", size=7) \
   .map(sns.distplot, "Age") \
   .add_legend();
plt.show()
```

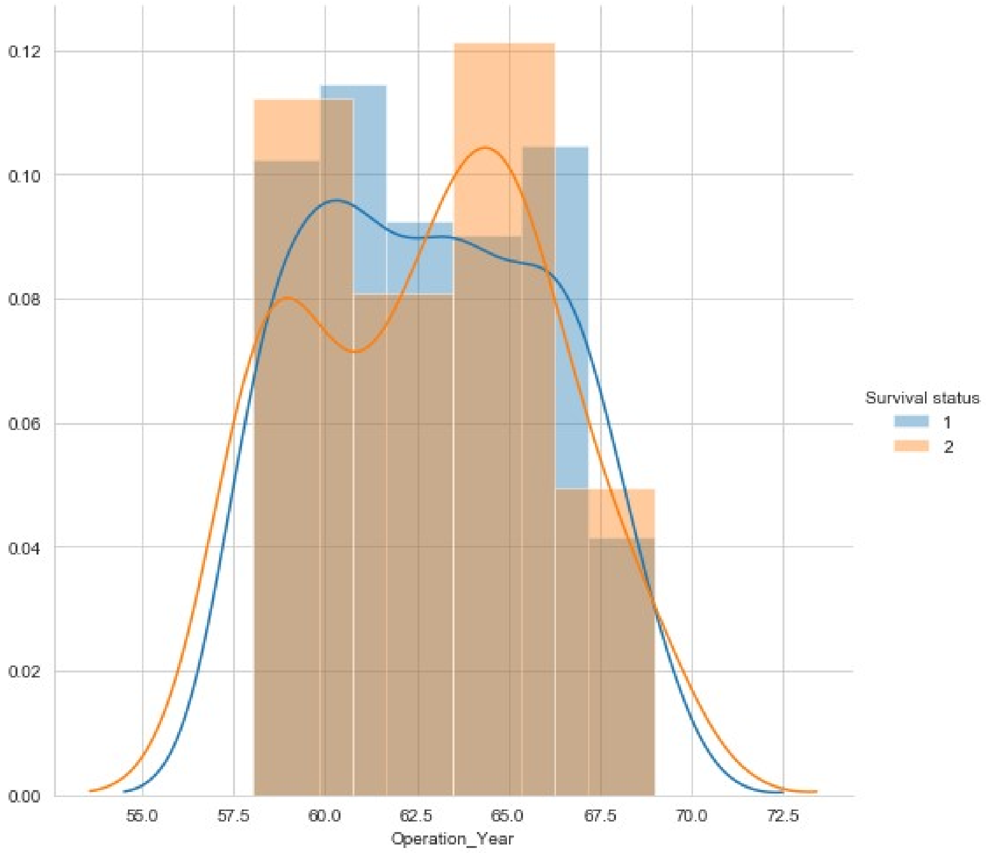
```
C:\Users\Sai charan\Anaconda3\lib\site-packages\seaborn\axisgrid.py:230: UserWar
ning: The `size` paramter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
C:\Users\Sai charan\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: Futur
eWarning: Using a non-tuple sequence for multidimensional indexing is deprecated
; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interp
reted as an array index, `arr[np.array(seq)]`, which will result either in an er
ror or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```



In [38]:

```
sns.FacetGrid(haberman, hue="Survival status", size=7) \
   .map(sns.distplot, "Operation_Year") \
   .add_legend();
plt.show();
```

```
C:\Users\Sai charan\Anaconda3\lib\site-packages\seaborn\axisgrid.py:230: UserWar
ning: The `size` paramter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
C:\Users\Sai charan\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: Futur
eWarning: Using a non-tuple sequence for multidimensional indexing is deprecated
; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interp
reted as an array index, `arr[np.array(seq)]`, which will result either in an er
ror or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```



Observations: 1.From the postive_lymph_nodes pdf distibution we can infer that most survival patients have fallen in to zero positive_lymph_nodes. 2.Age and survival status are not useful insights as the distibution is more similar for both people who survived and also dead. 3.people who didnt survive suddenly rise and fall in between 1958 and 1960. 4.more number of people are survived in the year 1965.

cdf

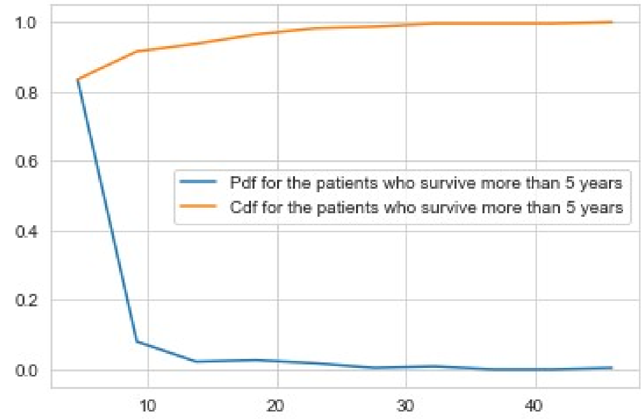
In [39]:

```
alive=haberman.loc[haberman["Survival status"]==1]
dead=haberman.loc[haberman["Survival status"]==2]
```

In [40]:

```
counts, bin_edges = np.histogram(alive['positive_lymph_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.legend(['Pdf for the patients who survive more than 5 years',
            'Cdf for the patients who survive more than 5 years'])
plt.show()
```

```
[0.83482143 0.08035714 0.02232143 0.02678571 0.01785714 0.00446429
 0.00892857 0.         0.         0.00446429]
[ 0.   4.6  9.2 13.8 18.4 23.  27.6 32.2 36.8 41.4 46. ]
```

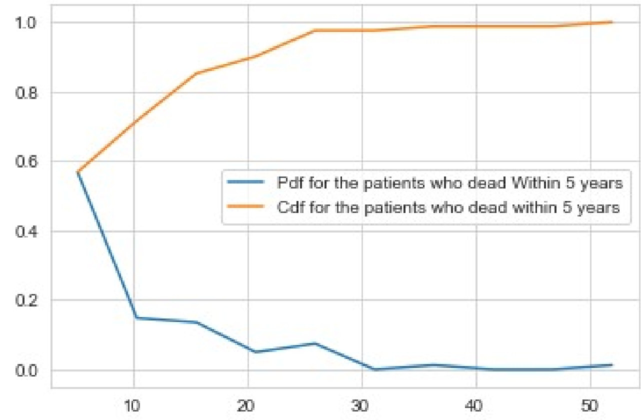


In [41]:

```
counts, bin_edges = np.histogram(dead['positive_lymph_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.legend(['Pdf for the patients who dead Within 5 years',
            'Cdf for the patients who dead within 5 years'])
plt.show()
```

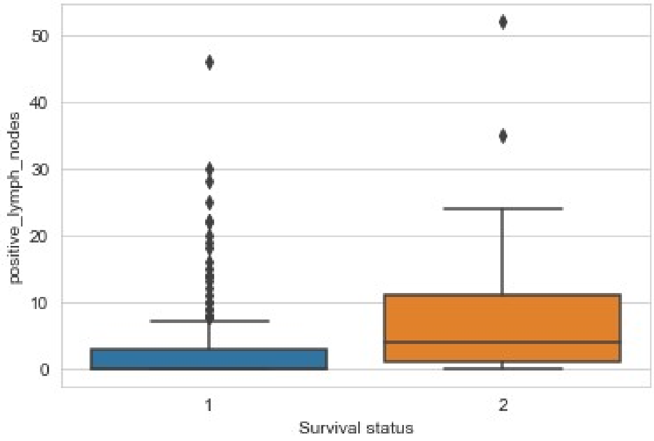
```
[0.56790123 0.14814815 0.13580247 0.04938272 0.07407407 0.
 0.01234568 0.         0.         0.01234568]
[ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
```



observations: 1.patients above 46 axillary nodes can be considered as dead within 5 years.

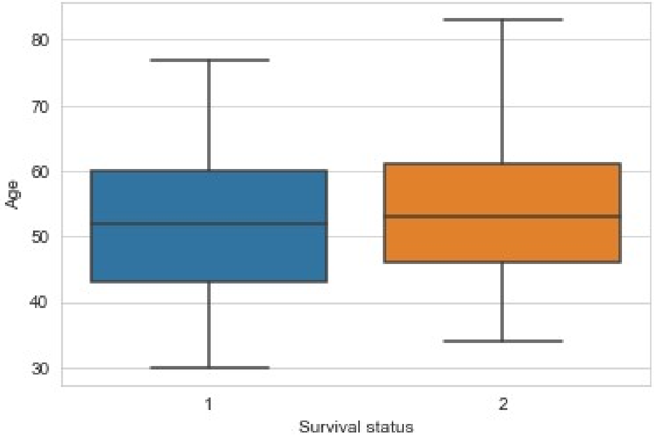
In [42]:

```
sns.boxplot(x='Survival status',y='positive_lymph_nodes', data=haberman)
plt.show()
```



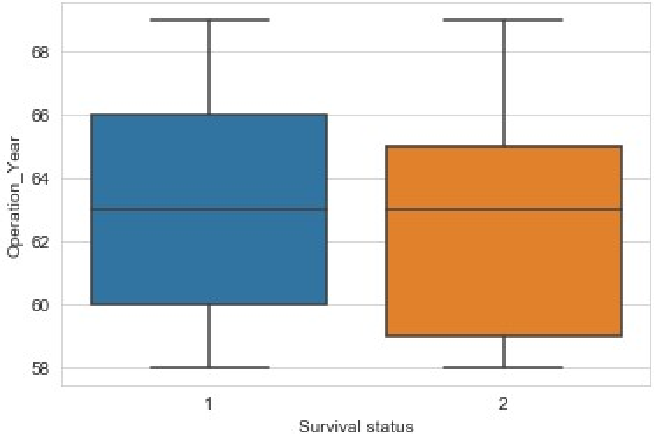
In [43]:

```
sns.boxplot(x='Survival status',y='Age', data=haberman)
plt.show()
```



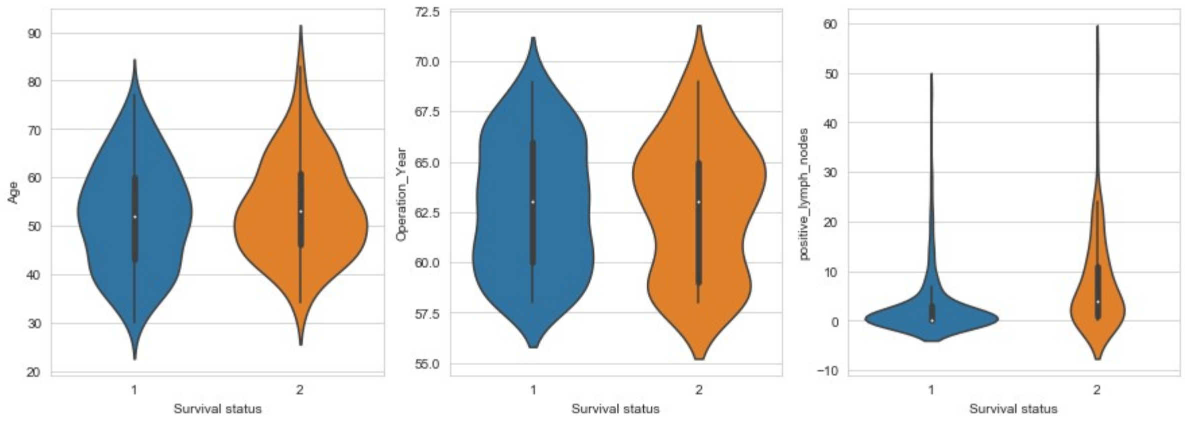
In [44]:

```
sns.boxplot(x='Survival status',y='Operation_Year', data=haberman)
plt.show()
```



In [45]:
```
fig, axes = plt.subplots(1, 3, figsize=(15, 5))
for idx, feature in enumerate(list(haberman.columns)[:-1]):
    sns.violinplot( x='Survival status', y=feature, data=haberman, ax=axes[idx])
plt.show()
```

```
C:\Users\Sai charan\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: Futur
eWarning: Using a non-tuple sequence for multidimensional indexing is deprecated
; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interp
reted as an array index, `arr[np.array(seq)]`, which will result either in an er
ror or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

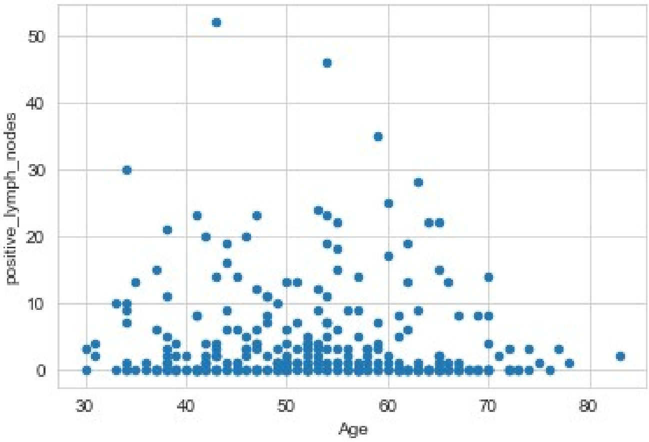


observations: 1.The density of number of positive lymph node is quite high between 0to 5. 2.number of patients who are dead have age between 46-62 than 59-65 and the patients who have survived are more in 42-60 than 60-66

BI-VARIATE ANALYSIS

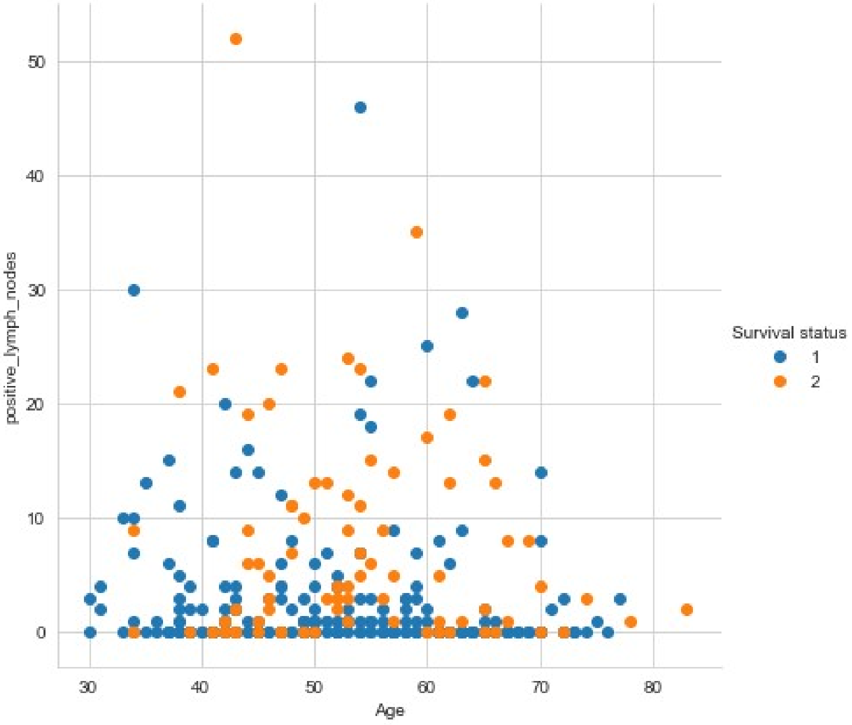
scatter plot

In [46]:
```
#
haberman.plot(kind='scatter', x='Age', y='positive_lymph_nodes') ;
plt.show()
```



In [47]:
```
sns.set_style("whitegrid");
sns.FacetGrid(haberman, hue="Survival status", size=6) \
   .map(plt.scatter, "Age", "positive_lymph_nodes") \
   .add_legend();
plt.show();
```

```
C:\Users\Sai charan\Anaconda3\lib\site-packages\seaborn\axisgrid.py:230: UserWar
ning: The `size` paramter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

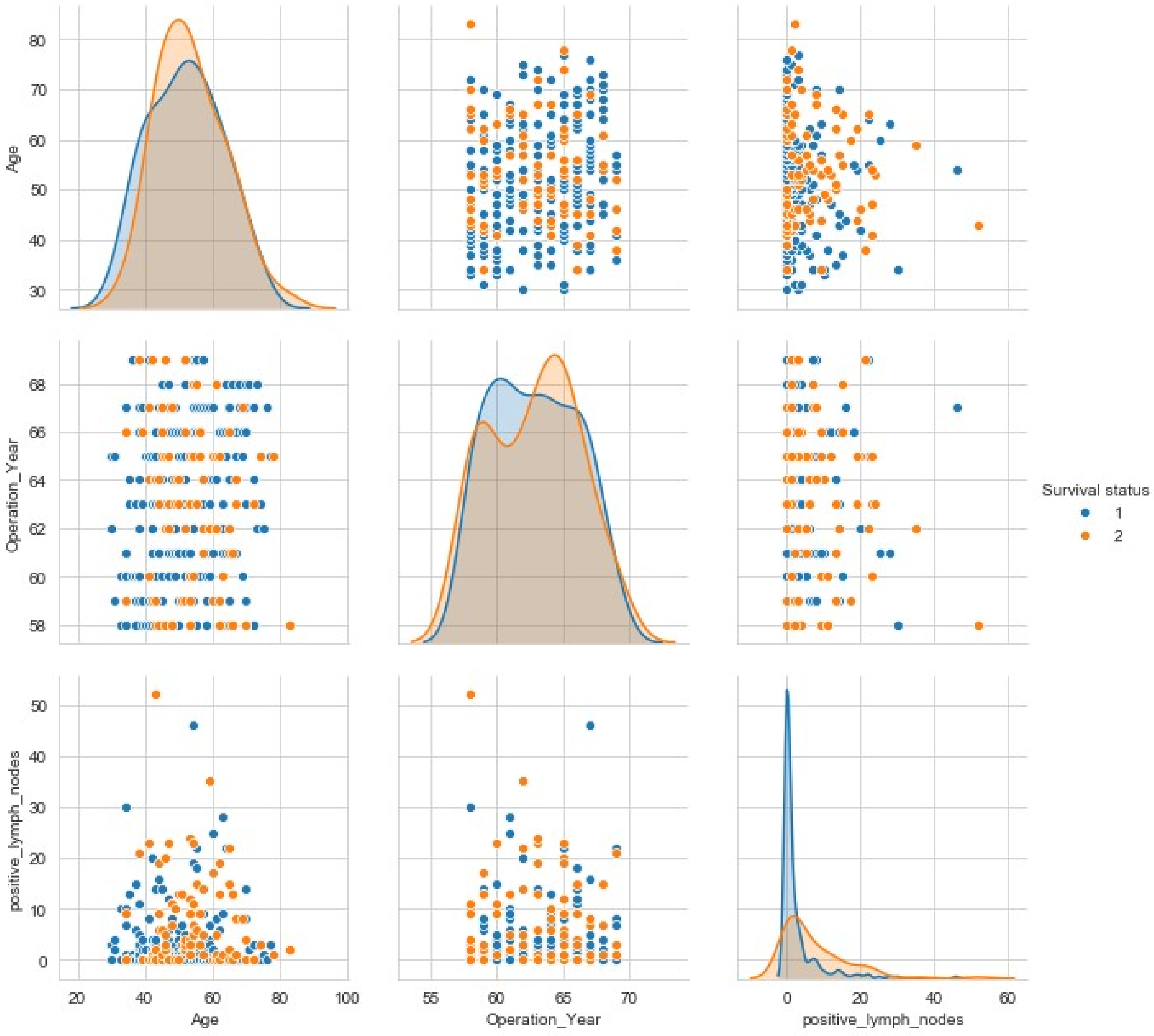


observations: 1.most of the patients have zero positive lymph nodes. 2.we cannot make any decision regarding patient's survival as the blue points are not seperated from orange points.

pair plot

```
In [48]: plt.close();
         sns.set_style("whitegrid");
         sns.pairplot(haberman, hue="Survival status",
                      vars=['Age','Operation_Year','positive_lymph_nodes'], size=3)
         plt.show()
```

```
C:\Users\Sai charan\Anaconda3\lib\site-packages\seaborn\axisgrid.py:2065: UserWa
rning: The `size` parameter has been renamed to `height`; pleaes update your cod
e.
  warnings.warn(msg, UserWarning)
C:\Users\Sai charan\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: Futur
eWarning: Using a non-tuple sequence for multidimensional indexing is deprecated
; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interp
reted as an array index, `arr[np.array(seq)]`, which will result either in an er
ror or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```



observations: 1.most people who survived have 0 postive lymph nodes detected by using positive lymph nodes vs age plot

Conclusions

1.There are 306 observations with 4 features in the data set

2.Uni-variate Analysis: a.pdF 1.From the postive_lymph_nodes pdf distibution we can infer that most survival patients have fallen in to 0 positive lymph nodes. 2.Age and survival status are not useful insights as the distibution is more similar for both people who survived and also dead. 3.people who didnt survive suddenly rise and fall in between 1958 and 1960. 4.more number of people are survived in the year 1965. b.cdf 1.patients above 46 axillary nodes can be considered as dead within 5 years. 3.box plot&violin plot 1.The density of number of positive lymph node is quite high between 0to 5. 2.number of patients who are dead have age between 46-62 than 59-65 and the patients who have survived are more in 42-60 than 60-66 3.Bi-variate Analysis: a.scatter plot 1.most of the patients have zero positive lymph nodes. 2.we cannot make any decision regarding patient's survival as the blue points are not seperated from orange points. b.Pair plot: 1.most people who survived have 0 postive lymph nodes detected by using positive lymph nodes vs age plot.