Plotting for Exploratory Data Analysis (EDA)

In [1]:

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

In [8]:

```
#loading data source in pandas dataframe
data=pd.read_csv("haberman.csv" , header=None ,names=['age','year','nodes','status'])
```

In [9]:

```
# data points and features
print(data.shape)
```

```
(306, 4)
```

In [10]:

```
print(data.columns)
```

```
Index(['age', 'year', 'nodes', 'status'], dtype='object')
```

In [11]:

```
data["status"].value_counts()
```

Out[11]:

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

Observation

In [12]:

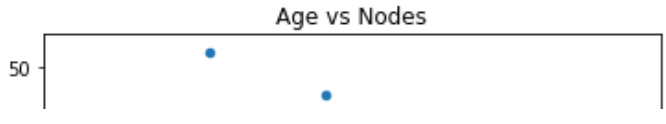
```
data.head()
```

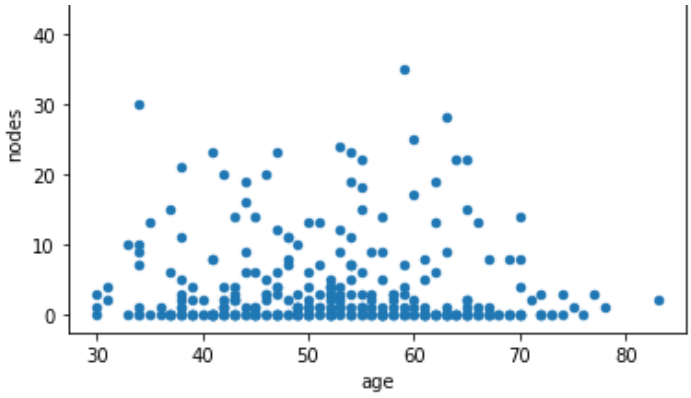
Out[12]:

	age	year	nodes	status
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1

In [14]:

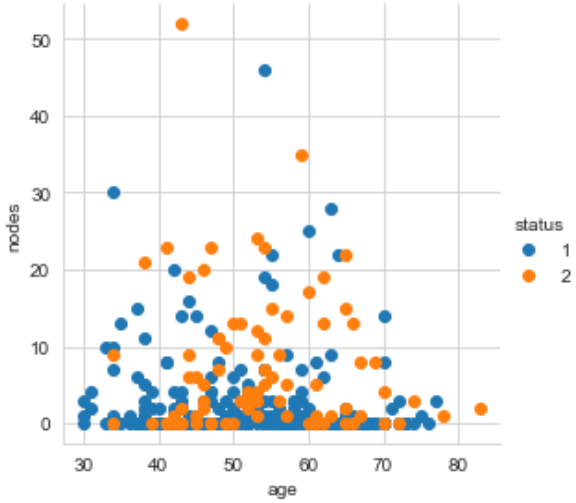
```
data.plot(kind='scatter', x='age',y='nodes')
plt.title("Age vs Nodes")
plt.show()
```





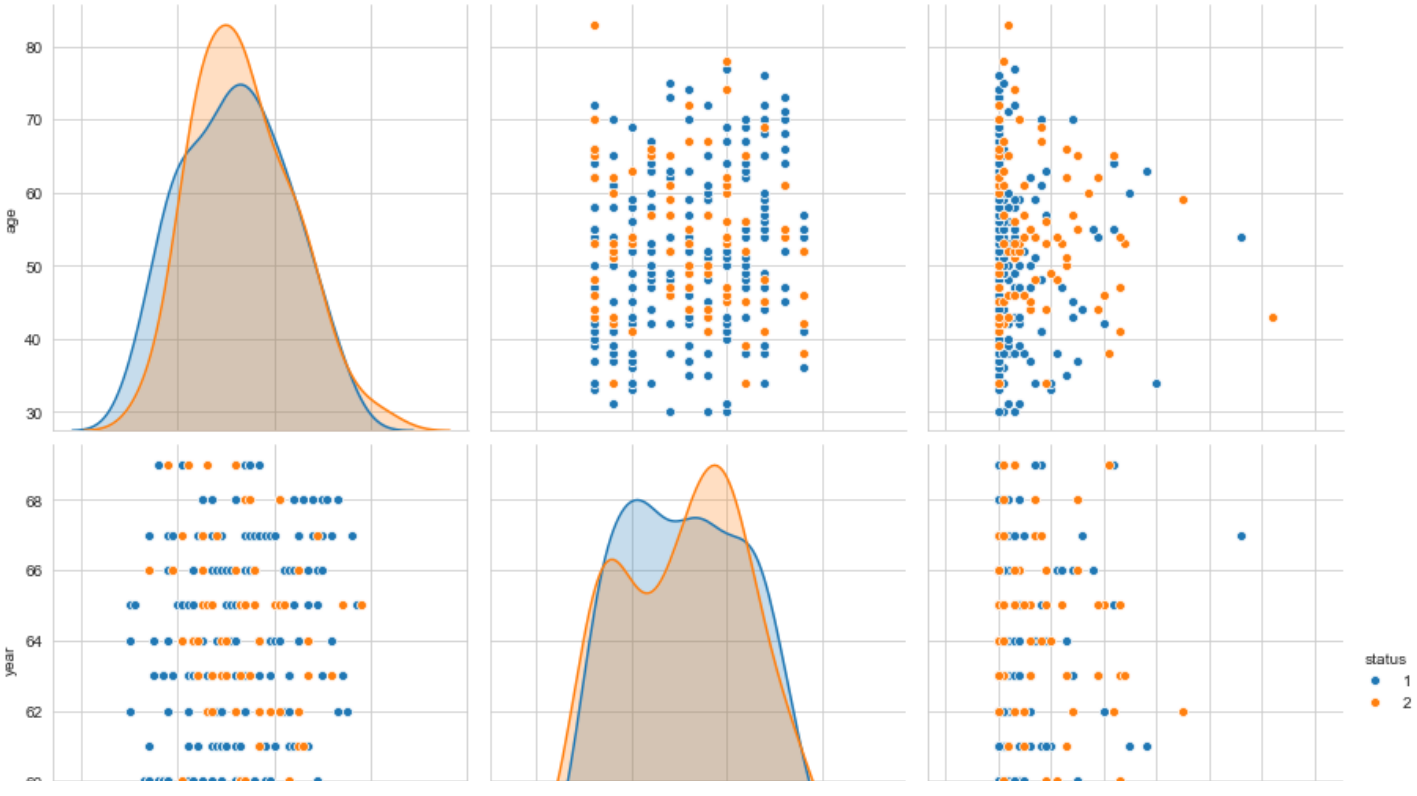
In [21]:

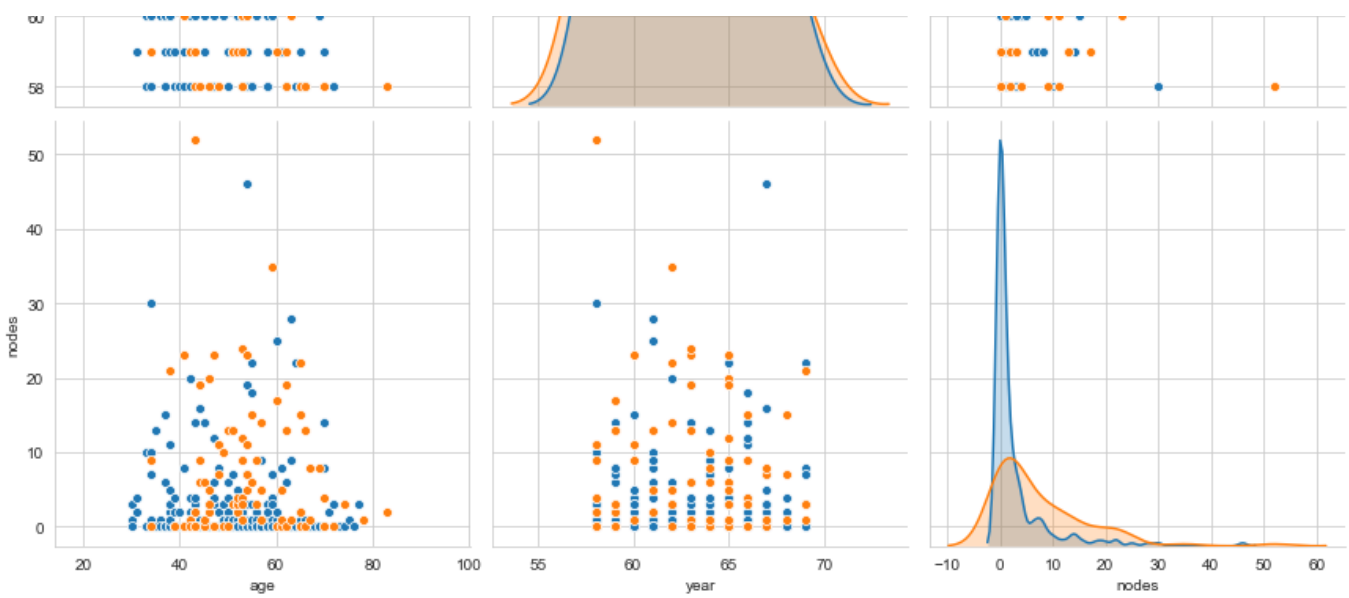
```
sns.set_style("whitegrid")
sns.FacetGrid(data,hue='status',size=4)\
    .map(plt.scatter ,"age","nodes")\
    .add_legend()
plt.show()
```



In [27]:

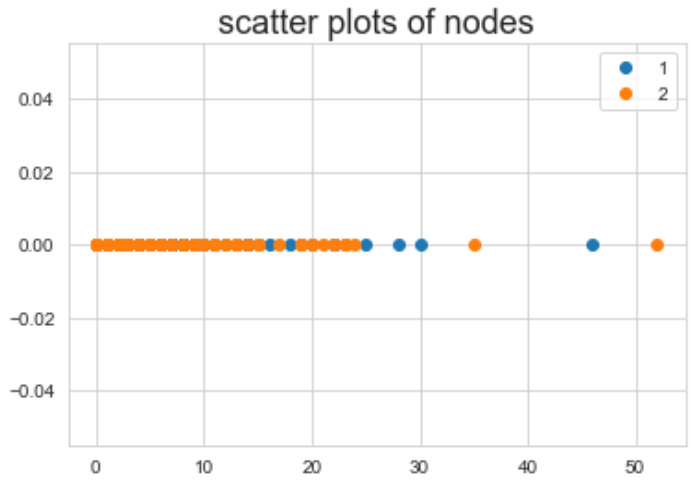
```
plt.close()
sns.set_style("whitegrid")
sns.pairplot(data,hue='status',height=4)
plt.show()
```





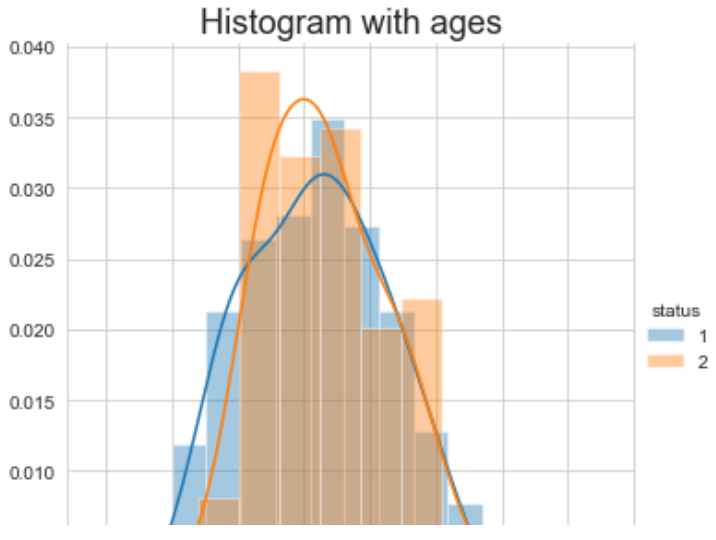
In [30]:

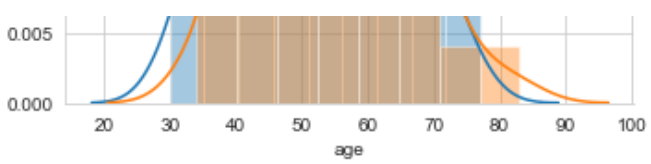
```
data_1=data.loc[data["status"]==1]
data_2=data.loc[data["status"]==2]
plt.plot(data_1["nodes"],np.zeros_like(data_1["nodes"]),'o',label="1")
plt.plot(data_2["nodes"],np.zeros_like(data_2["nodes"]),'o',label="2")
plt.title("scatter plots of nodes",size=18)
plt.legend()
plt.show()
```



In [41]:

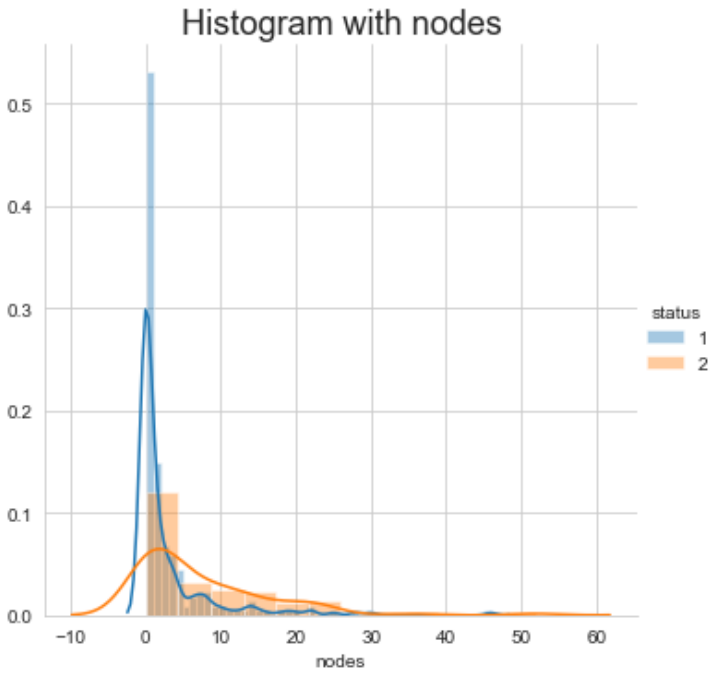
```
sns.FacetGrid(data , hue="status" , height=5)\
.map(sns.distplot,"age")\
.add_legend();
plt.title("Histogram with ages",size=18)
plt.show()
```





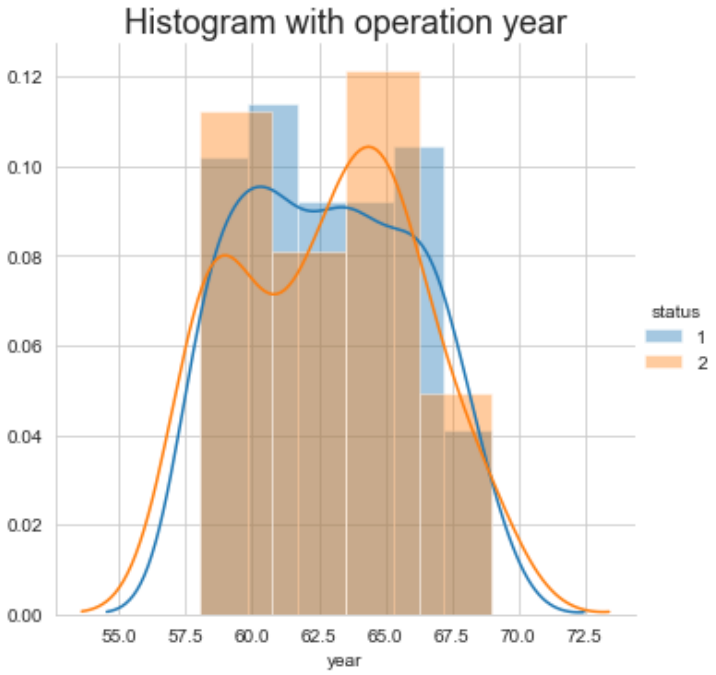
In [42]:

```
sns.FacetGrid(data , hue="status" , height=5)\
.map(sns.distplot,"nodes")\
.add_legend();
plt.title("Histogram with nodes",size=18)
plt.show()
```



In [43]:

```
sns.FacetGrid(data , hue="status" , height=5)\
.map(sns.distplot,"year")\
.add_legend();
plt.title("Histogram with operation year",size=18)
plt.show()
```

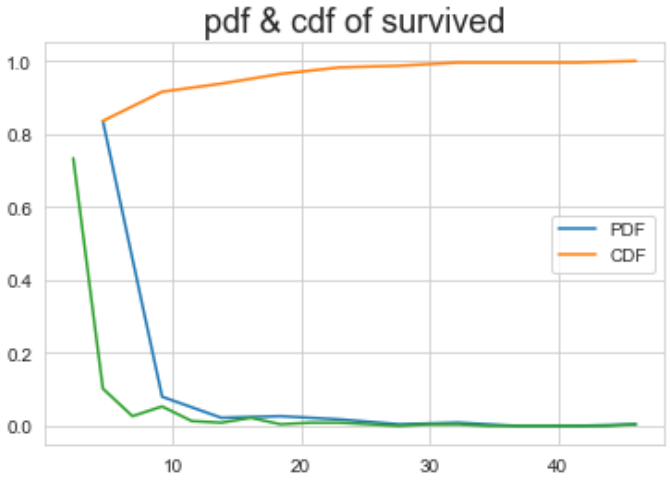


In [44]:

```
counts,bin_edges=np.histogram(data_1['nodes'],bins=10,density=True)
```

```
pdf=counts/sum(counts)
print(pdf)
print(bin_edges)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label="PDF")
plt.plot(bin_edges[1:],cdf,label="CDF")

counts,bin_edges=np.histogram(data_1['nodes'],bins=20,density=True)
pdf=counts/sum(counts)
plt.plot(bin_edges[1:],pdf)
plt.legend(loc="best")
plt.title("pdf & cdf of survived" , size=18)
plt.show()
```

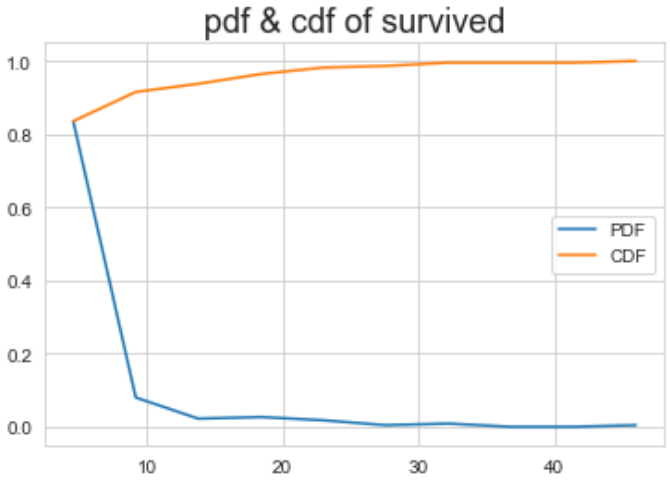
```
[0.83555556 0.08       0.02222222 0.02666667 0.01777778 0.00444444
 0.00888889 0.         0.         0.00444444]
[ 0.   4.6  9.2 13.8 18.4 23.  27.6 32.2 36.8 41.4 46. ]
```



In [45]:

```
counts,bin_edges=np.histogram(data_1['nodes'],bins=10,density=True)
pdf=counts/sum(counts)
print(pdf)
print(bin_edges)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label="PDF")
plt.plot(bin_edges[1:],cdf,label="CDF")
plt.title("pdf & cdf of survived" , size=18)
plt.legend()
plt.show()
```

```
[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. ]
```

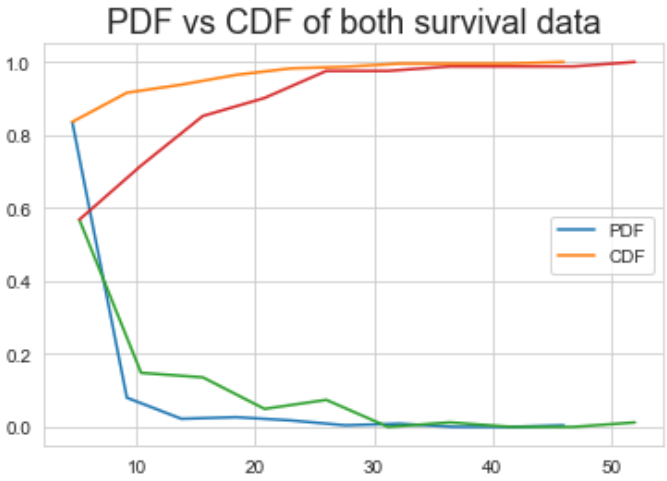


In [48]:

```
counts,bin_edges=np.histogram(data_1['nodes'],bins=10,density=True)
```

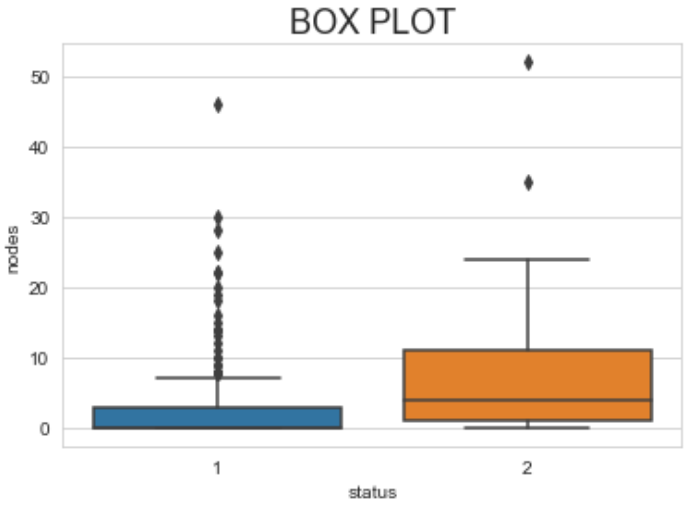
```
pdf=counts/sum(counts)
print(pdf)
print(bin_edges)
cdf=np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf,label="PDF")
plt.plot(bin_edges[1:],cdf,label="CDF")

counts,bin_edges=np.histogram(data_2['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()
plt.title("PDF vs CDF of both survival data" , size=18)
plt.show()
```

```
[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. ]
[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. ]
```



In [50]:

```
sns.boxplot(x='status', y='nodes',data=data)
plt.title("BOX PLOT",size=18)
plt.show()
```



CONCLUSION

- **At first i want to describe about the haberman dataset which contains of 304 instances and features are 4 which is 1.age of patient 2. operation year 3. axillary nodes 4. survival status**

- **With the help of features we tried to know the relations between features of data**
- **For knowing the relations between data, we visualize the data with the help of diffrent typs of plot**
- **In this analysis of haberman data i did univarate analysis and multivariate analysis**
- **As we can say, less number of data are available abode node 30**
- **After doing all the visualisation and analysis, we can conclude many things as follows such as we can say most of operations done one age range between 58 and 66 with the help of scatter plot**
- **And large amount of data are available at axil node 0, with scatter plot we can say that at nodes 0 are more chances to survive**
- **We can also say that the more number of nodes you have more chances to die**
- **And the nodes features are the most important feature to know the status of patient whether he has more chances to die or not.**
- **Patients having age less than 40 are less chances to die**
- **With the help of box plot, we can say that most of the person who survived at node 0.**