In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
#load data
haberman=pd.read_csv("C:/Users/VINAY PRATAP SINGH/Desktop/haberman.csv"
)
```

In [3]:
```
haberman.head()
```

Out[3]:

	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 [4]:
```
#Number of data point
print(haberman.shape)
```

```
(306, 4)
```

In [5]:
```
#number of features
print(haberman.columns)#it gives column name and data type
```

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

In [6]:
```
#number of classes
haberman['status'].value_counts()
```

```
#there are 2 class:-
# 1 :-the patient survived after operation
# 2 :-the patient died after operation
```

Out[6]:
```
1    225
2     81
Name: status, dtype: int64
```

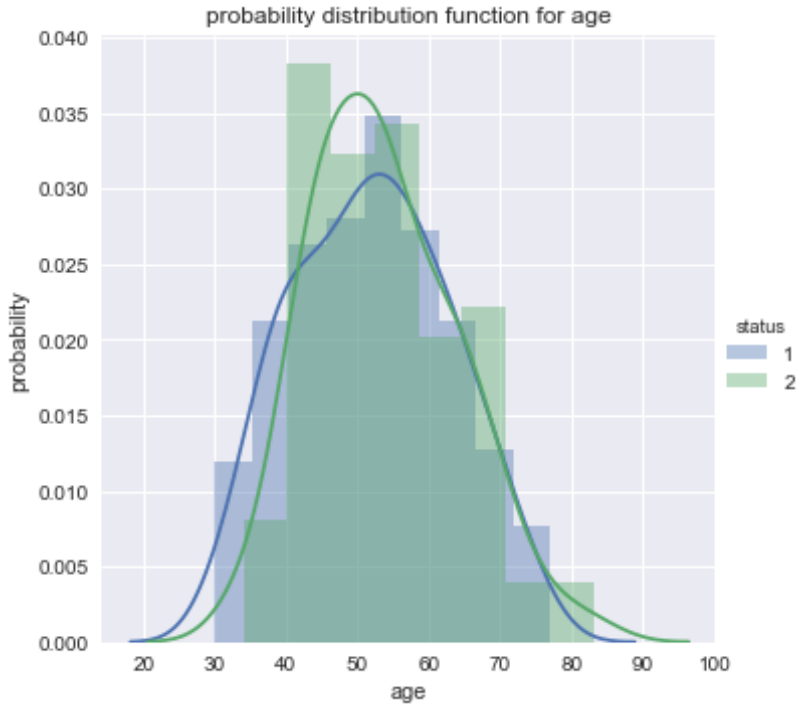
Objective :-

Our obective is to classify the patient who are survived after the operation and the patient who are dead after the operation.

Univaraite analysis(PDF, CDF, Boxplot, Voilin plots)

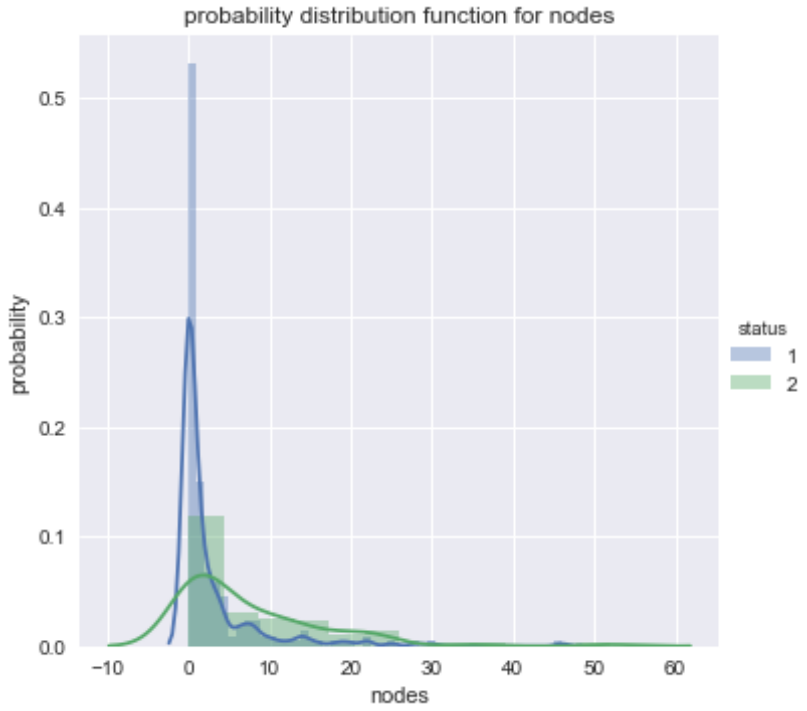
In [12]:
```
#PDF:-Probability distribution function
#plotting 1-D
sns.FacetGrid(haberman, hue='status', size=5) \
    .map(sns.distplot, 'age') \
    .add_legend()
plt.ylabel("probability")
plt.title("probability distribution function for age")

plt.show()
```



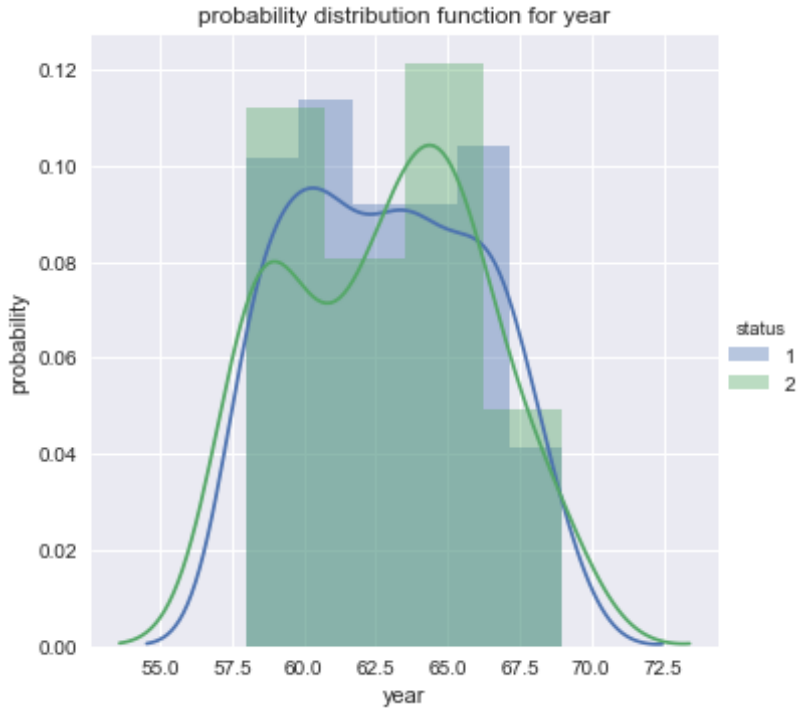
In [13]:
```
sns.FacetGrid(haberman, hue='status', size=5) \
    .map(sns.distplot, 'nodes') \
    .add_legend()
plt.ylabel("probability")
plt.title("probability distribution function for nodes")

plt.show()
```



In [14]:
```
sns.FacetGrid(haberman, hue='status', size=5) \
    .map(sns.distplot, 'year') \
    .add_legend()
plt.ylabel("probability")
plt.title("probability distribution function for year")

plt.show()
```



Observation

In above figures features are overlapping to each other. But when we plot the feature "node" it shows that node <=5, the patient is survived else dead.

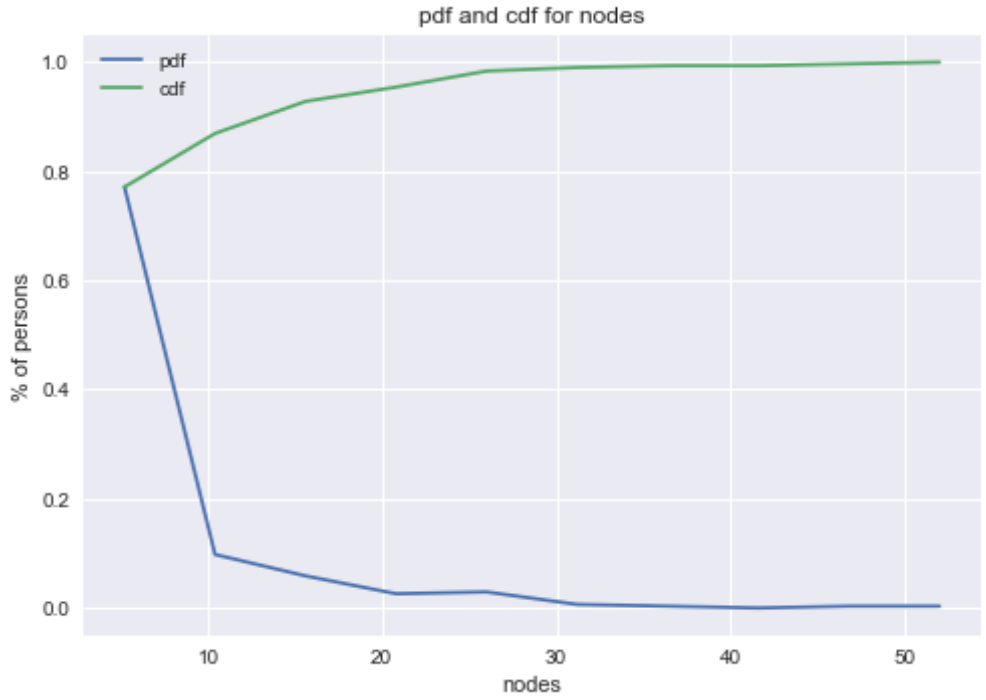
CDF

In [19]:
```
#CDF:-cumulative density function
#1-D plot

counts, bin_edges = np.histogram(haberman['nodes'], bins=10, density=True)
legend=["pdf","cdf"]
```

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


plt.xlabel("nodes")
plt.ylabel("% of persons")
plt.title("pdf and cdf for nodes")
plt.legend(legend)
plt.show()
```

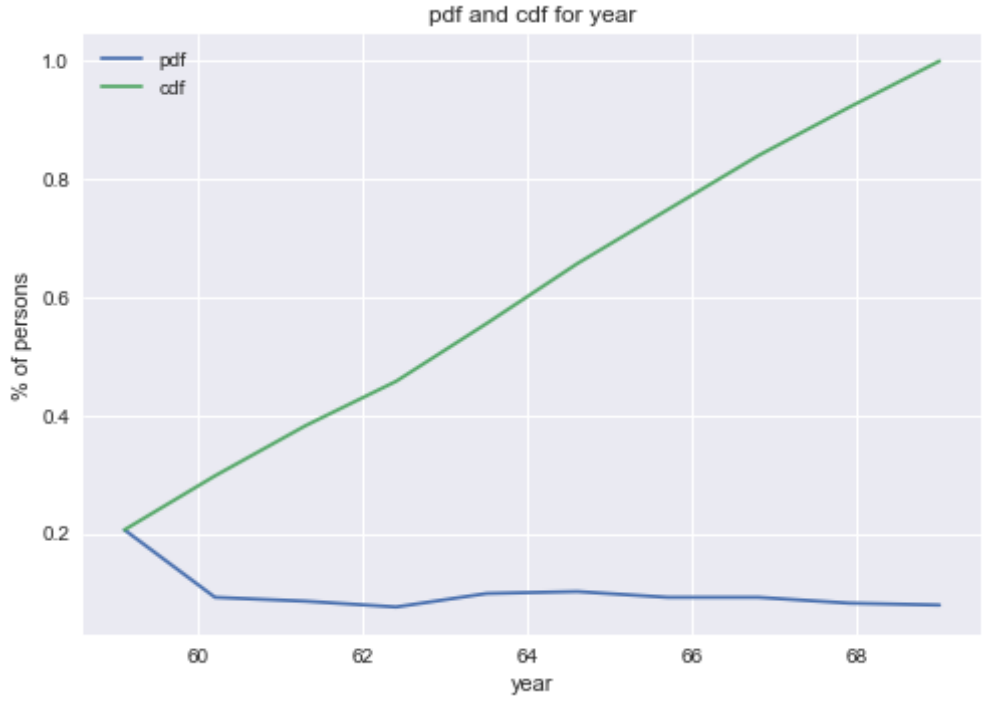


Observation

From the above graph we can say that 100% had less than 40 nodes and 40% have less than 10 nodes.

In [20]:
```
counts, bin_edges = np.histogram(haberman['year'], bins=10, density=True)
legend=["pdf","cdf"]
pdf = counts/sum(counts)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:], pdf)
plt.plot(bin_edges[1:], cdf)


plt.xlabel("year")
plt.ylabel("% of persons")
plt.title("pdf and cdf for year")
plt.legend(legend)
plt.show()
```

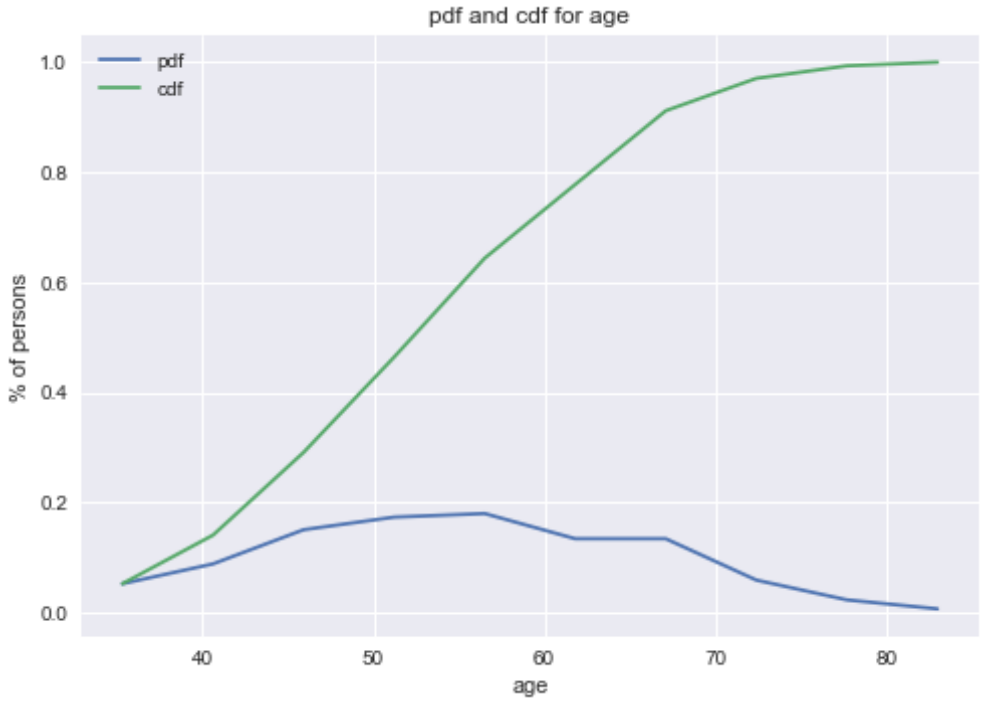


Observation

From above graph we can say that 95% people have birth year below 68 and 20% people have birth year below 60.

In [21]:
```
counts, bin_edges = np.histogram(haberman['age'], bins=10, density=True
)
legend=["pdf","cdf"]
pdf = counts/sum(counts)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:], pdf)
plt.plot(bin_edges[1:], cdf)


plt.xlabel("age")
plt.ylabel("% of persons")
plt.title("pdf and cdf for age")
plt.legend(legend)
plt.show()
```

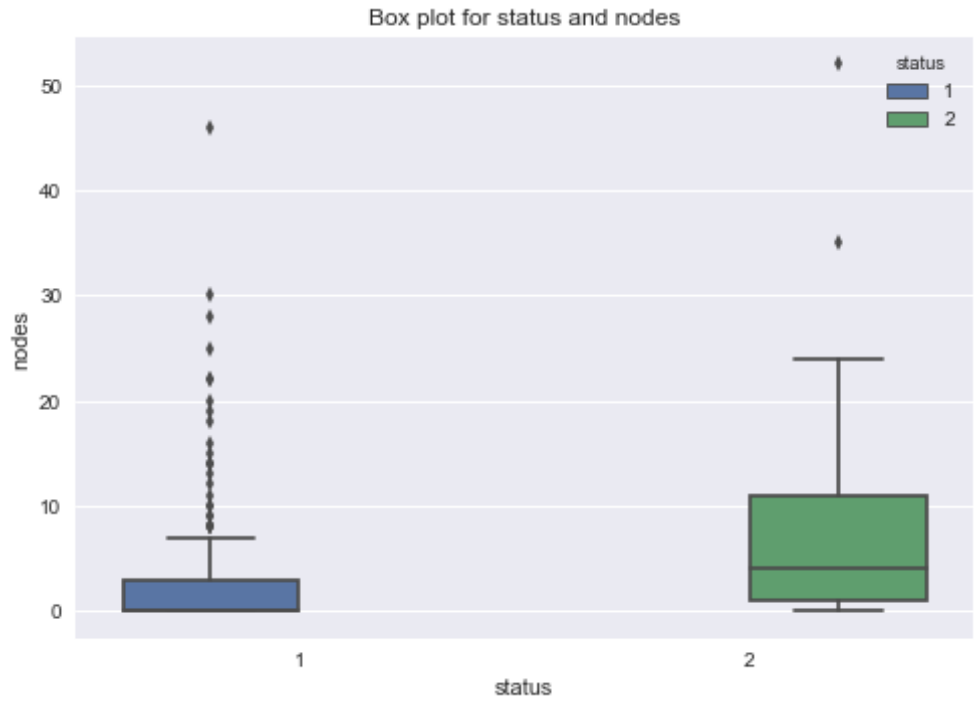


Observation

From above graph we can say that 100% people have age below 80year and 5% people have age below 40year.

Boxplot

In [24]:
```
sns.boxplot(x = "status", y = "nodes", hue = "status", data = haberman)
.set_title("Box plot for status and nodes")
plt.show()
```

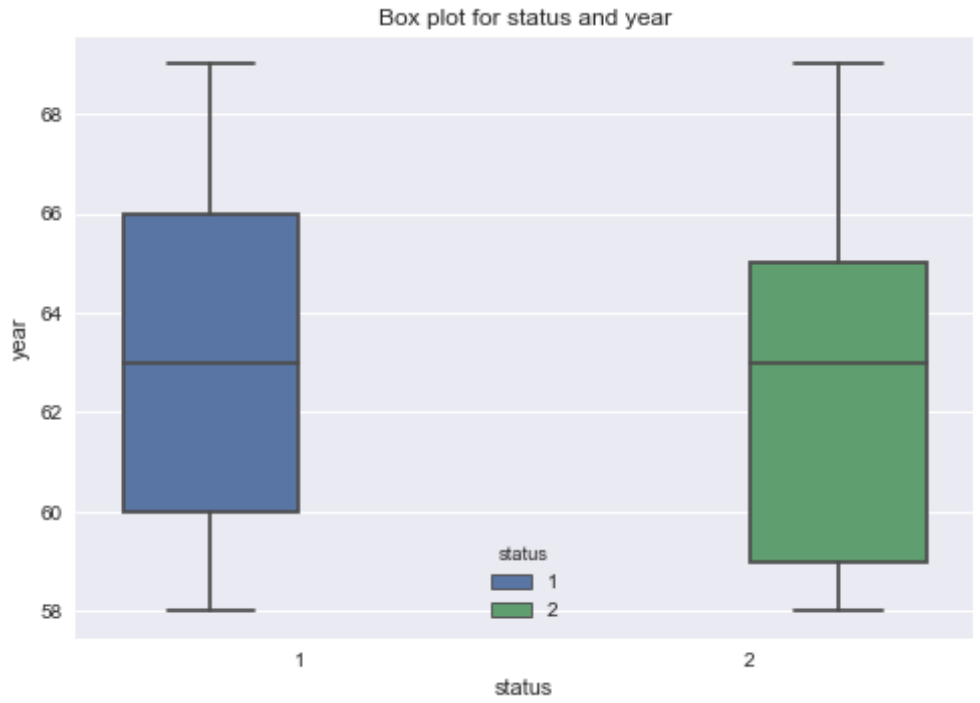


Observation

At boxplot 1, the survive people on 75th percentile have 2 nodes, 25th and 50th percentiles are overlapped and above whiskers outliers are present .

At boxplot 2, the dead people on 25th percentile have 1 node, 50th percentile have 3 nodes, 75th percentile have 11 nodes above whiskers outliers are present.

In [23]:
```
sns.boxplot(x = "status", y = "year", hue = "status", data = haberman).
set_title("Box plot for status and year")
plt.show()
```

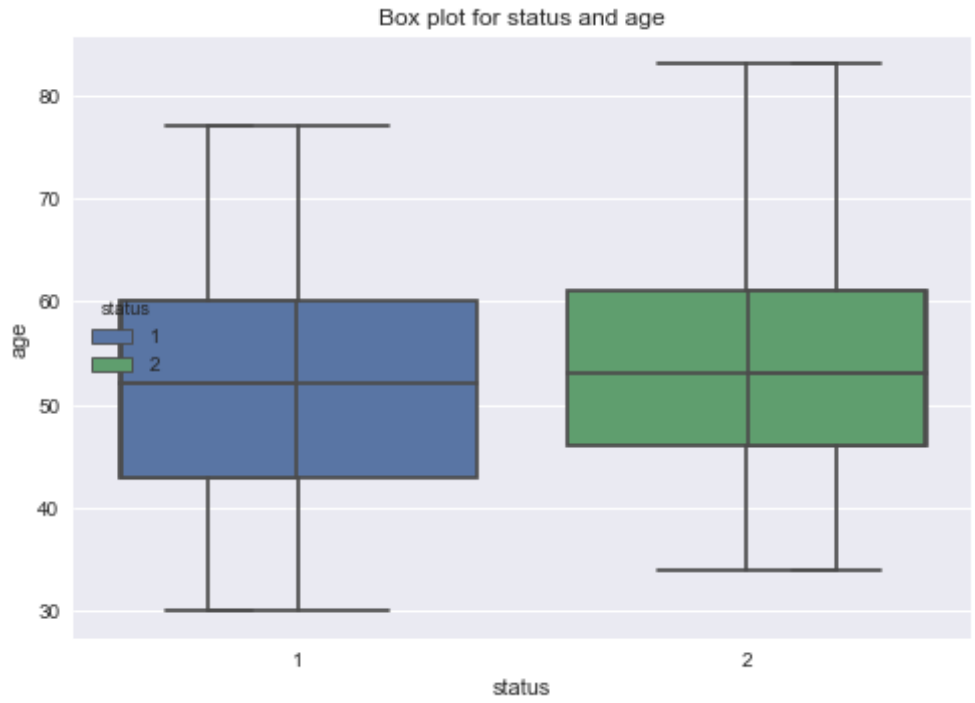


observation

At boxplot 1, the survive people year on 25th percentile have year 60, 50th percentile have year 63 and 75th prcentile have year 66.

At boxplot 2,the dead people year on 25th percentile have year 59, 50th percentile have year 63 and 75th percentile have year 65.

In [29]:
```
sns.boxplot(x = "status", y = "age",hue="status", data = haberman).set_
title("Box plot for status and age")
plt.show()
```



Observation

At boxplot 1,the survive people age on 25th percentile have age 42, 50th percentile have age 52, 75th perentile have age 60.

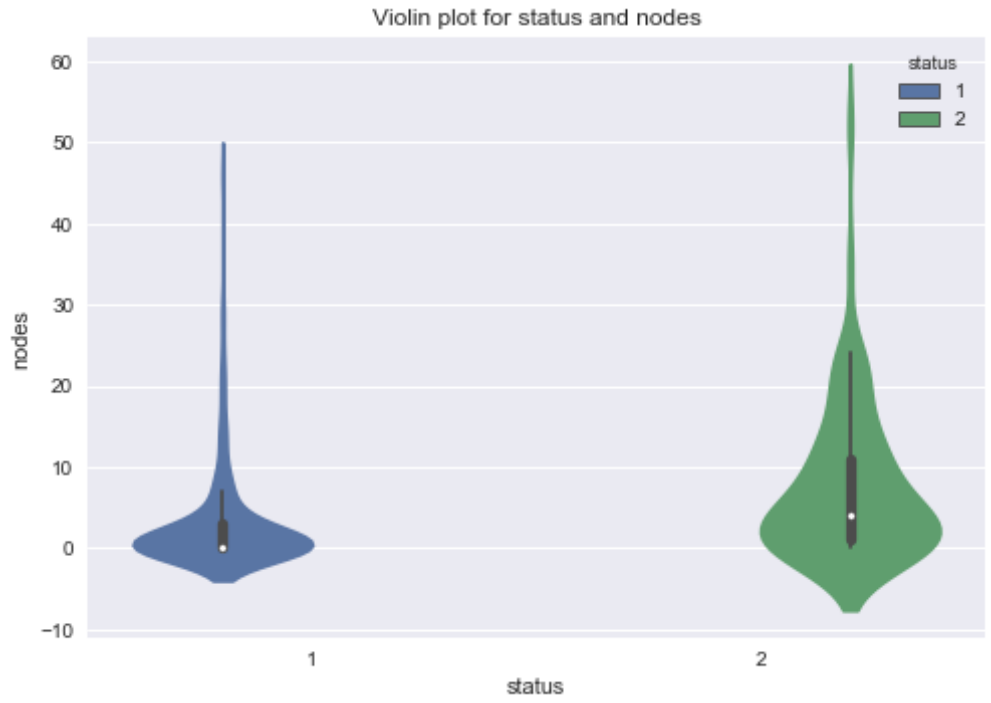
At boxplot 2,the dead people age on 25th percentile have age 46, 50th percentile have age 53, 75th percentile have age 61.

Voilin plots

In [30]:
```
#It is a graph in which the histogram and boxplot are represent

sns.violinplot(x = "status", y = "nodes", hue = "status", data = haberm
```

```
an)
plt.title("Violin plot for status and nodes")
plt.show()
```



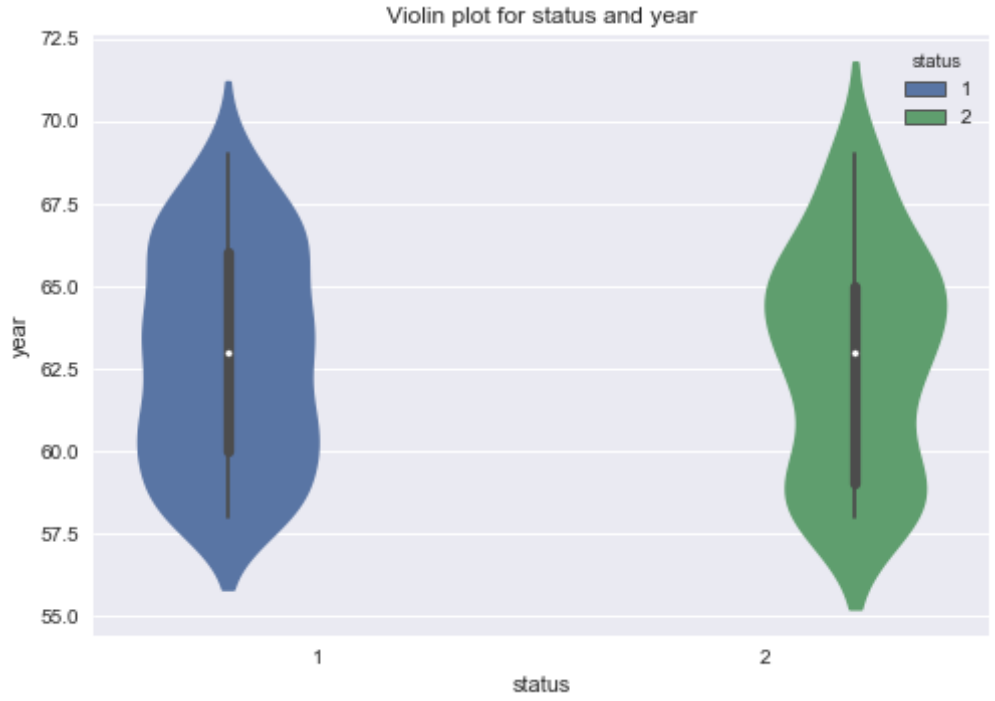
Observation

At violin plot 1, 50th percentile of survive person have 0 nodes, 75th percentile have less than 3 positive nodes.

At violin plot 2, 25th percentile of dead have 1 node, 50th percentile of dead have nodes below 4,75th percentile of dead have nodes below 11.

In [31]:
```
sns.violinplot(x = "status", y = "year", hue = "status", data = haberma
n)
```

```
plt.title("Violin plot for status and year")
plt.show()
```

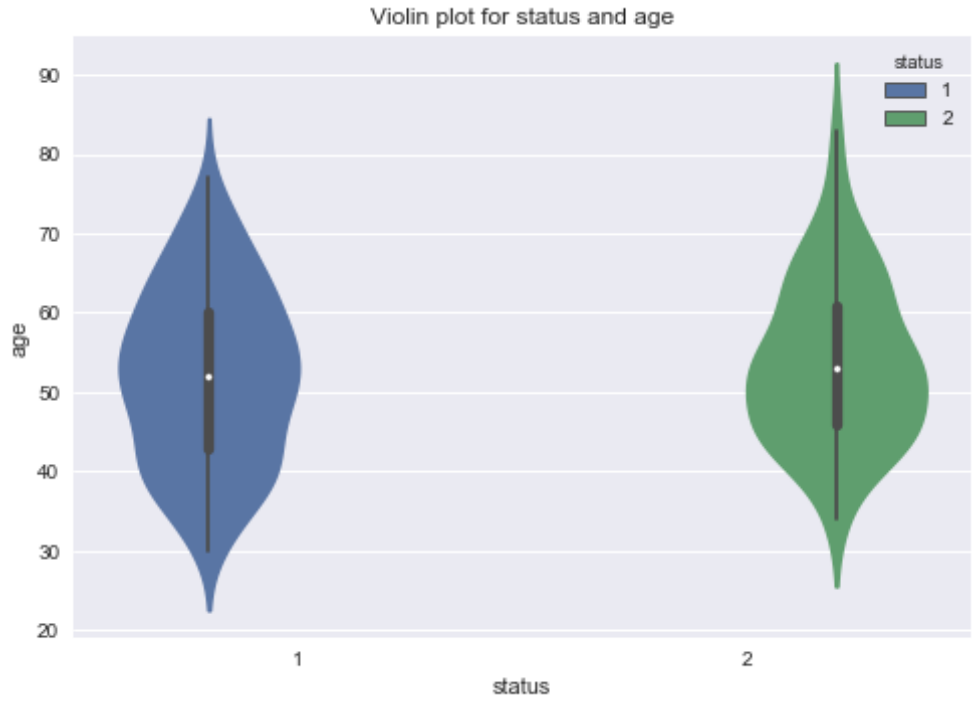


Observation

At violin plot 1, 25th percentile of survive person have year 60, 50th percentile of survive person have year 63 and 75th percentile have year 66.

At violin plot 2, 25th percentile of dead have year 58.5, 50th percentile of dead have year 63,75th percentile of dead have year 65.

In [32]:
```
sns.violinplot(x = "status", y = "age", hue = "status", data = haberman
)
plt.title("Violin plot for status and age")
plt.show()
```



Observation

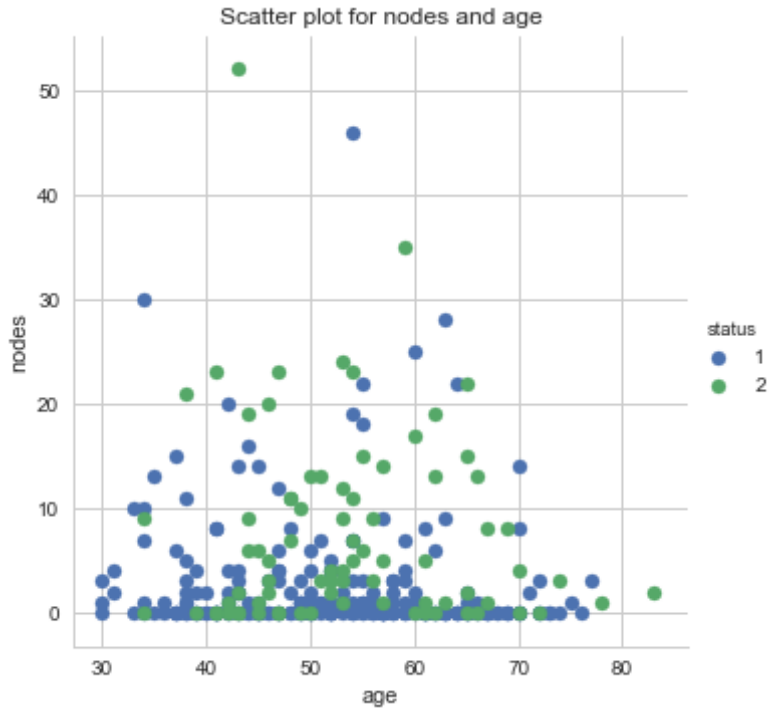
At violin plot 1, 25th percentile of survive person have age 42, 50th percentile of survive person have age 52 and 75th percentile have age 60.

At violin plot 2, 25th percentile of dead have age 46, 50th percentile of dead have year 53,75th percentile of dead have year 61.

Bi-variate analysis (scatter plots, pair-plots)

Scatter plots

In [33]:
```
# For Age and  nodes
sns.set_style('whitegrid')
sns.FacetGrid(haberman, hue='status', size=5) \
    .map(plt.scatter, 'age', 'nodes') \
    .add_legend()
plt.title("Scatter plot for nodes and age")
plt.show()
```



Observation:

In above scatter plot graph shows that 1 and 2 status overlapped to each other using "age" and "nodes" features so the classifiaction between both is not possible.

In [34]:
```
# For Age and Year
```

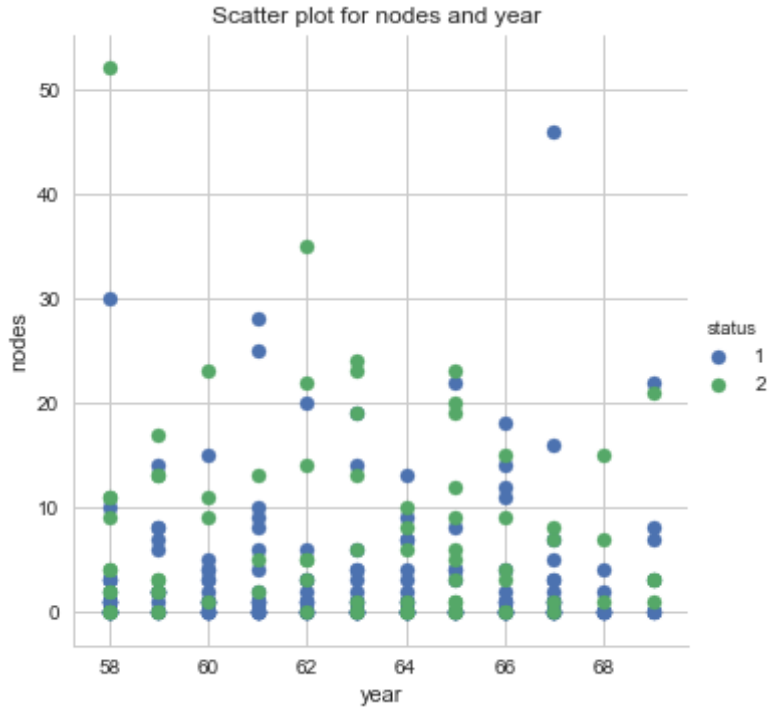
```
sns.set_style('whitegrid')
sns.FacetGrid(haberman, hue="status", size=5) \
    .map(plt.scatter, 'age', 'year') \
    .add_legend()
plt.title("Scatter plot for year and age")
plt.show()
```

Observation

Classification is not possible because of overlapping occur between two variables.

In [35]:
```
# For Year and  nodes
sns.set_style('whitegrid')
sns.FacetGrid(haberman, hue='status', size=5) \
    .map(plt.scatter, 'year', 'nodes') \
```

```
    .add_legend()
plt.title("Scatter plot for nodes and year ")
plt.show()
```



Observation

Classification is not possible because of overlapping.

Pair plot

In [51]:
```
sns.set_style('whitegrid')
sns.pairplot(haberman, hue='status', vars=['age', 'year', 'nodes'], siz
```

```
e=5)
plt.show()
```

Observation

Classification is not possible because of overlapping between variables.

Conclusion

By plotting all pdf, cdf, box-plot, violin plot, pair plots, scatter plot etc. we get only one conclusion : if number of node is less,than survival of patients is more.

In [ ]: