HabermanDataset: The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

General info to get a clear view:

- The person typically has around 20-30 axillary lymph nodes(lymph nodes in the armpits).
- After a breast cancer diagnosis, a doctor may check whether cancer cells have spread to the axillary lymph nodes. This can help confirm the diagnosis and staging of cancer.

Source:
1.https://www.kaggle.com/datasets/gilsousa/habermans-survival-data-set
2.https://www.medicalnewstoday.com/articles/319713#:~:text=Axillary%20lymph%20nodes%20are%20located,20%E2%80%9330%20axillary%20lymph%20nodes.

Objective: The objective is to Analyse the DataSet,and to gather the information through EDA.

In [1]:
```
#importing the required libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
df=pd.read_csv('haberman.csv') # Reading the csv file

df.columns      #Checking the column names of DataSet.
```

Out[2]:
```
Index(['30', '64', '1', '1.1'], dtype='object')
```

Given DataSet doesn't have labels(column names).

In [3]:
```
#Naming the labels referenced from the source of the given dataset.

df.columns=['age','year','nodes','status']
```

- age: Petiont's Age
- year: Operation Year(1958-1970)
- nodes: infected Axillary nodes detected in petiont.
- status: Whether petiont has Survived or Not Survived

In [4]:
```
#Replacing Variables for better understanding
df.status[df['status']==1]='Survived'
df.status[df['status']==2]='NotSurvived'
```

In [5]:
```
df.head()              #overview of DataSet
```

Out[5]:

	age	year	nodes	status
0	30	62	3	Survived
1	30	65	0	Survived
2	31	59	2	Survived
3	31	65	4	Survived
4	33	58	10	Survived

In [6]:
```
#Description of DataSet
print(df.describe())
print("\nshape ",df.shape,"\n")
print(df.status.value_counts(),"\n\n...")
```

```
              age        year       nodes
count  305.000000  305.000000  305.000000
mean    52.531148   62.849180    4.036066
std     10.744024    3.254078    7.199370
min     30.000000   58.000000    0.000000
25%     44.000000   60.000000    0.000000
50%     52.000000   63.000000    1.000000
75%     61.000000   66.000000    4.000000
max     83.000000   69.000000   52.000000

shape  (305, 4)

Survived       224
NotSurvived     81
Name: status, dtype: int64

...
```

By observing the records in given DataSet

- The average age of petionts is **52 years**.
- The Youngest Petiont is of **30 years old**,whereas Eldest is **83 years old**.
- The survival rate of petionts is about **73%** (224/305*100=73.44%)

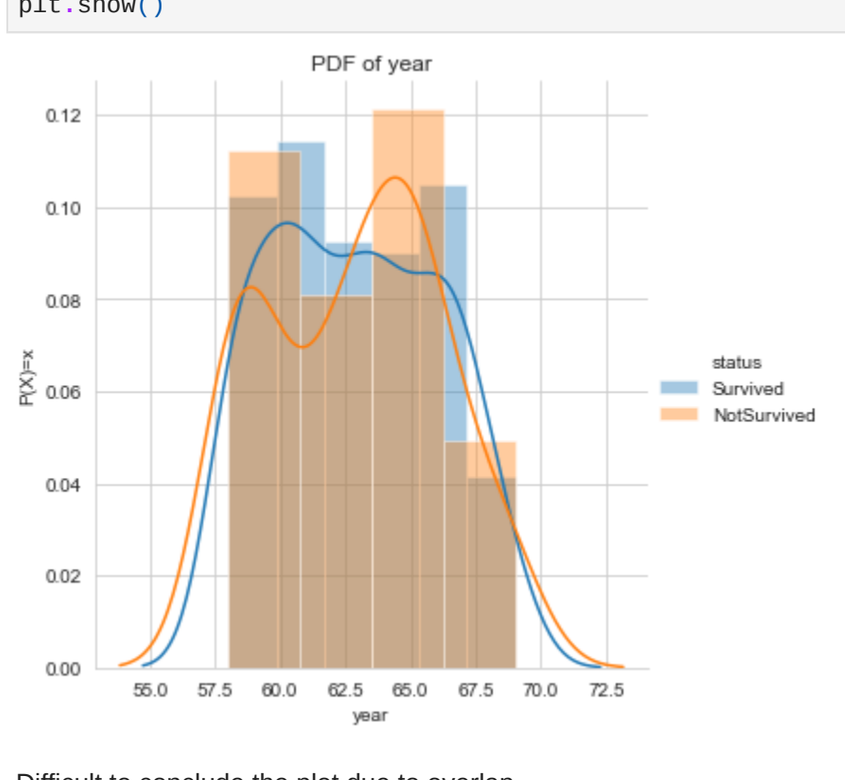
According to above observation the diffrence between above classes is found to be higher. Hence,our **DataSet is imbalanced**.

Univariate Analysis

PDF

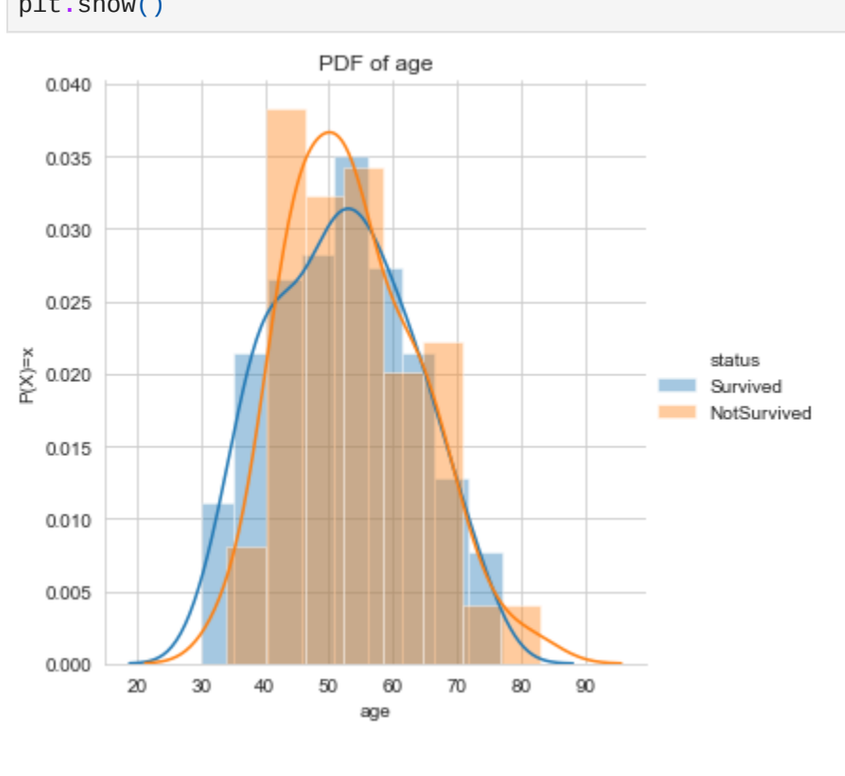
*If a Function follows continuous Destribution,The PDF(probability Density function) is a probability that a random variable **X** will take value exactly equal to **x**(a particular value)*
In simpler words, PDF is a percetage of data at that point.

In [20]:
```
sns.FacetGrid(df,hue='status',size=5).map(sns.distplot,'year').add_legend()
plt.ylabel('P(X)=x')
plt.title('PDF of year')
plt.show()
```

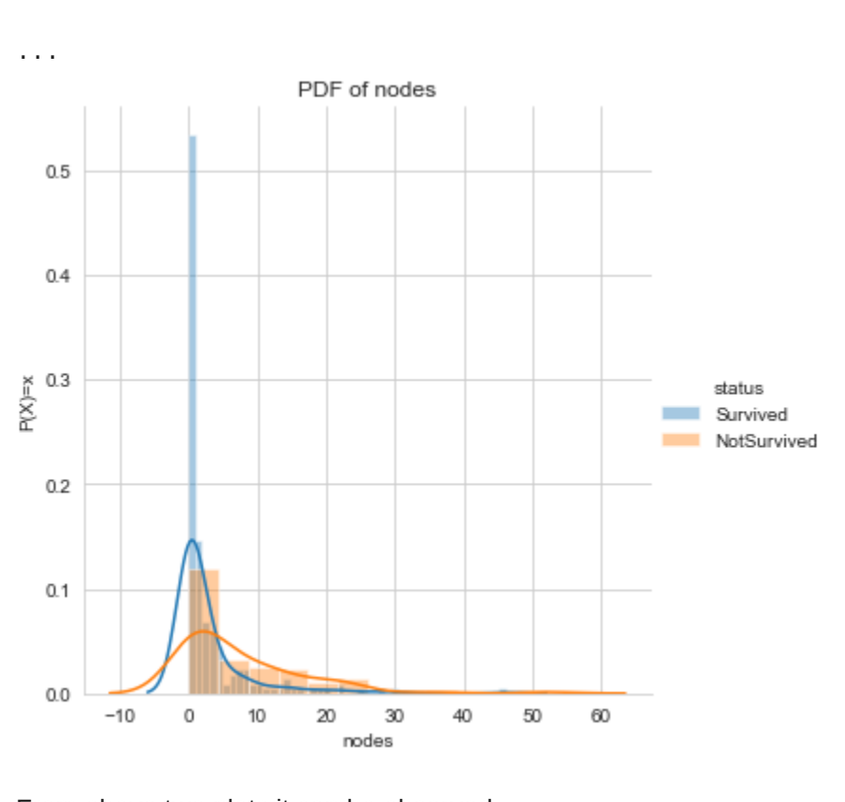


Difficult to conclude the plot due to overlap.

In [19]:
```
sns.FacetGrid(df,hue='status',size=5).map(sns.distplot,'age').add_legend()
plt.title('PDF of age')
plt.ylabel('P(X)=x')
plt.show()
print("\n\n...")
sns.FacetGrid(df,hue='status',size=5).map(sns.distplot,'nodes').add_legend()
plt.ylabel('P(X)=x')
plt.title('PDF of nodes')
plt.show()
```



...



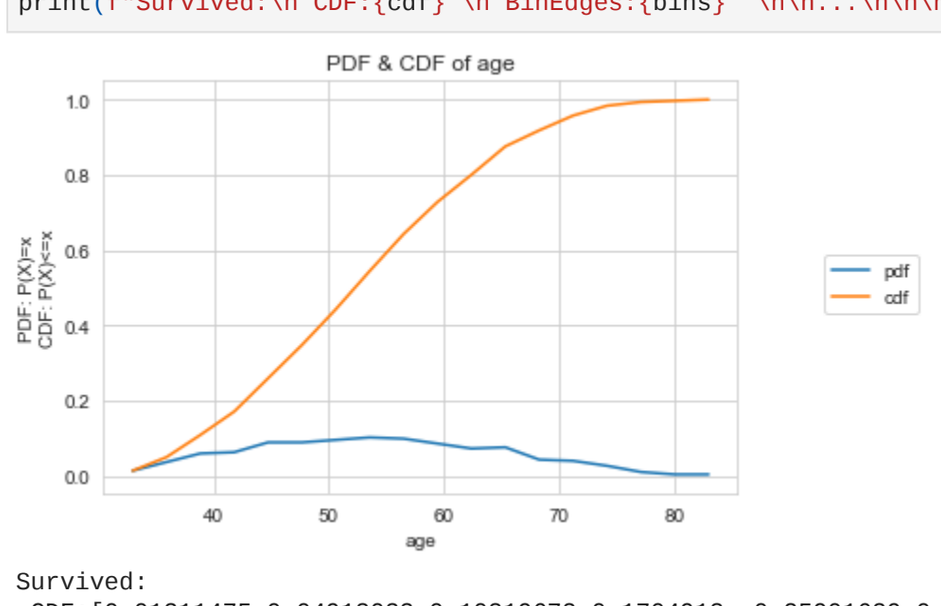
From above two plots it can be observed

- 1. people having age **less than or equal to 39 have higher chances of survival** ,wheras it gets inversed with people aged greater than 39.*(statement is unclear)*
- 1. lesser the cancer spread across nodes **higher the chances survival**,the statement has found out to be correct(As most of the plot have verified or concluded our statement.)
- 1. Chances are sligtly higher that the person will not survive if cancer has spread to **more than 4 nodes**.

As,more the conclusion come out of Data better it is...we will try getting more conclusions by plotting few more plots.

CDF
*The Cumulative Destribution Function is the probability that a random variable **X** will take value lesser than equals to **x**(perticular value).*
In simpler words,CDF tells the percetage of data lesser than equal to that point(x).

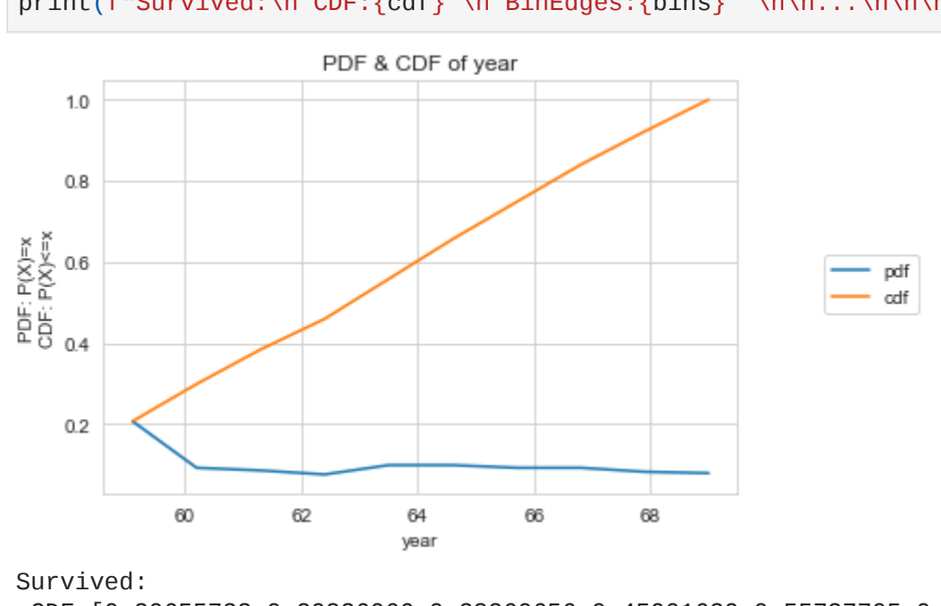
In [104…
```
counts,bins=np.histogram(df['age'],bins=18,density=True)
pdf=counts/(sum(counts))
cdf=np.cumsum(pdf)


plt.plot(bins[1:],pdf,label='pdf')
plt.plot(bins[1:],cdf,label='cdf')

plt.xlabel("age")
plt.ylabel("PDF: P(X)=x\nCDF: P(X)<=x")
plt.legend(bbox_to_anchor=(1.3, 0.6))
plt.title("PDF & CDF of age")
plt.show()
print(f"Survived:\n CDF:{cdf} \n BinEdges:{bins}  \n\n...\n\n\n\n")
```




```
Survived:
 CDF:[0.01311475 0.04918033 0.10819672 0.1704918  0.25901639 0.34754098
 0.44262295 0.5442623  0.64262295 0.72786885 0.8        0.87540984
 0.91803279 0.95737705 0.98360656 0.99344262 0.99672131 1.        ]
 BinEdges:[30.         32.94444444 35.88888889 38.83333333 41.77777778 44.72222222
 47.66666667 50.61111111 53.55555556 56.5        59.44444444 62.38888889
 65.33333333 68.27777778 71.22222222 74.16666667 77.11111111 80.05555556
 83.        ]

...
```

Around 1% of petionts have age lesser than 30 years.
Around 44% of petionts have age lesser than 50 years.
Around 80% of petionts have age lesser than 71 years.
Around 99% of petionts have age lesser than 80 years.

In [111…
```
counts,bins=np.histogram(df['year'],bins=10,density=True)
pdf=counts/(sum(counts))
cdf=np.cumsum(pdf)


plt.plot(bins[1:],pdf,label='pdf')
plt.plot(bins[1:],cdf,label='cdf')

plt.xlabel("year")
plt.ylabel("PDF: P(X)=x\nCDF: P(X)<=x")
plt.legend(bbox_to_anchor=(1.3, 0.6))
plt.title("PDF & CDF of year")
plt.show()
print(f"Survived:\n CDF:{cdf} \n BinEdges:{bins}  \n\n...\n\n\n\n")
```

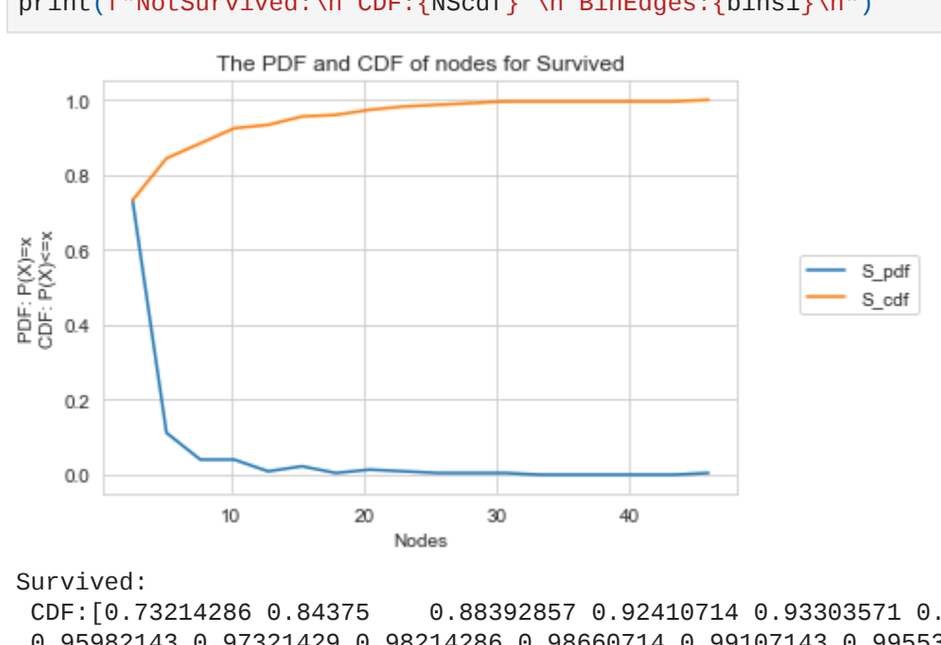



```
Survived:
 CDF:[0.20655738 0.29836066 0.38360656 0.45901639 0.55737705 0.6557377
 0.74754098 0.83934426 0.92131148 1.        ]
 BinEdges:[58.  59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]

...
```

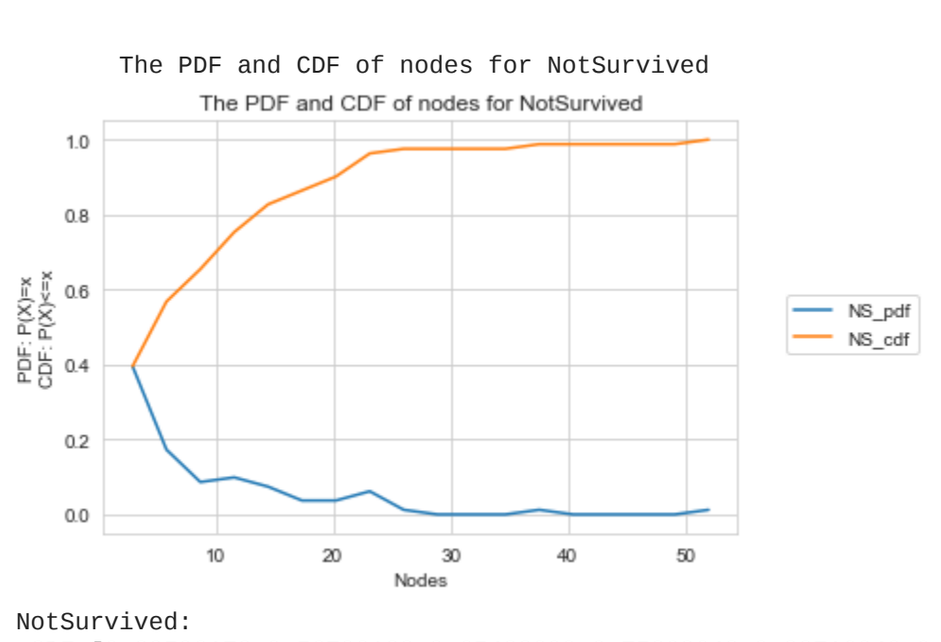
Around 55% of petionts had been operated year lesser than or equal to 63.

Due to the limited number of operation years in DataSet,We are unable to draw many conclusions.

In [72]:
```
Survived=df['nodes'][df['status']=='Survived']
NotSurvived=df['nodes'][df['status']=='NotSurvived']

counts,bins=np.histogram(Survived,bins=18,density=True)
pdf=counts/(sum(counts))
cdf=np.cumsum(pdf)


plt.plot(bins[1:],pdf,label='S_pdf')
plt.plot(bins[1:],cdf,label='S_cdf')

plt.xlabel("Nodes")
plt.ylabel("PDF: P(X)=x\nCDF: P(X)<=x")
plt.legend(bbox_to_anchor=(1.3, 0.6))
plt.title("The PDF and CDF of nodes for Survived")
plt.show()
print(f"Survived:\n CDF:{cdf} \n BinEdges:{bins}  \n\n...\n\n\n\n")


print("\n\n       The PDF and CDF of nodes for NotSurvived")
counts1,bins1=np.histogram(NotSurvived,bins=18,density=True)
NSpdf=counts1/(sum(counts1))
NScdf=np.cumsum(NSpdf)

plt.plot(bins1[1:],NSpdf,label='NS_pdf')
plt.plot(bins1[1:],NScdf,label='NS_cdf')
plt.xlabel("Nodes")
plt.ylabel("PDF: P(X)=x\nCDF: P(X)<=x")
plt.legend(bbox_to_anchor=(1.3, 0.6))
plt.title("The PDF and CDF of nodes for NotSurvived")
plt.show()
print(f"NotSurvived:\n CDF:{NScdf} \n BinEdges:{bins1}\n")
```




```
Survived:
 CDF:[0.73214286 0.84375    0.88392857 0.92410714 0.93303571 0.95535714
 0.95982143 0.97321429 0.98214286 0.98660714 0.99107143 0.99553571
 0.99553571 0.99553571 0.99553571 0.99553571 0.99553571 1.        ]
 BinEdges:[ 0.          2.55555556  5.11111111  7.66666667 10.22222222 12.77777778
 15.33333333 17.88888889 20.44444444 23.         25.55555556 28.11111111
 30.66666667 33.22222222 35.77777778 38.33333333 40.88888889 43.44444444
 46.        ]

...



       The PDF and CDF of nodes for NotSurvived
```




```
NotSurvived:
 CDF:[0.39506173 0.56790123 0.65432099 0.75308642 0.82716049 0.86419753
 0.90123457 0.96296296 0.97530864 0.97530864 0.97530864 0.97530864
 0.98765432 0.98765432 0.98765432 0.98765432 0.98765432 1.        ]
 BinEdges:[ 0.          2.88888889  5.77777778  8.66666667 11.55555556 14.44444444
 17.33333333 20.22222222 23.11111111 26.         28.88888889 31.77777778
 34.66666667 37.55555556 40.44444444 43.33333333 46.22222222 49.11111111
 52.        ]
```

In [ ]:
```
#Values are evaluated by slightly changing bin counts.
```

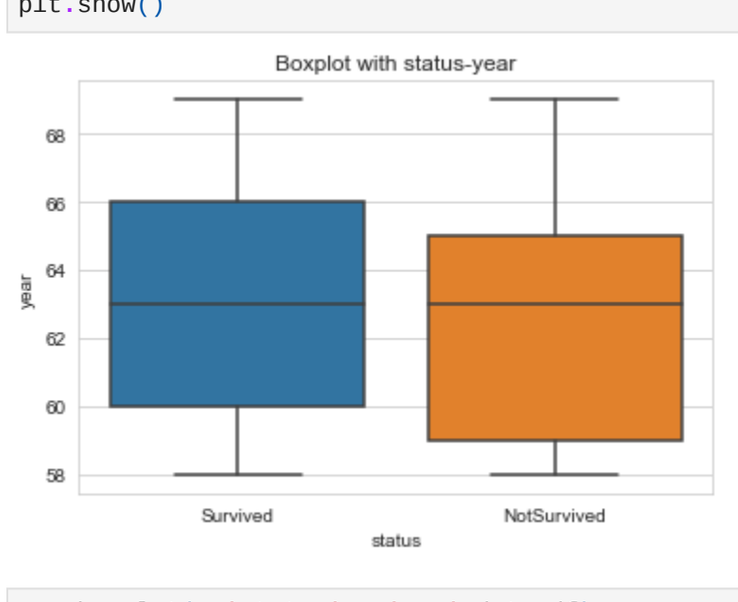
Tabular representation:

Nodes	Survived	NotSurvived	Difference
2	73%	39%	34%
5	84%	56%	28%
10	92%	71%	21%
15	95%	85%	10%
30	99%	97%	3%

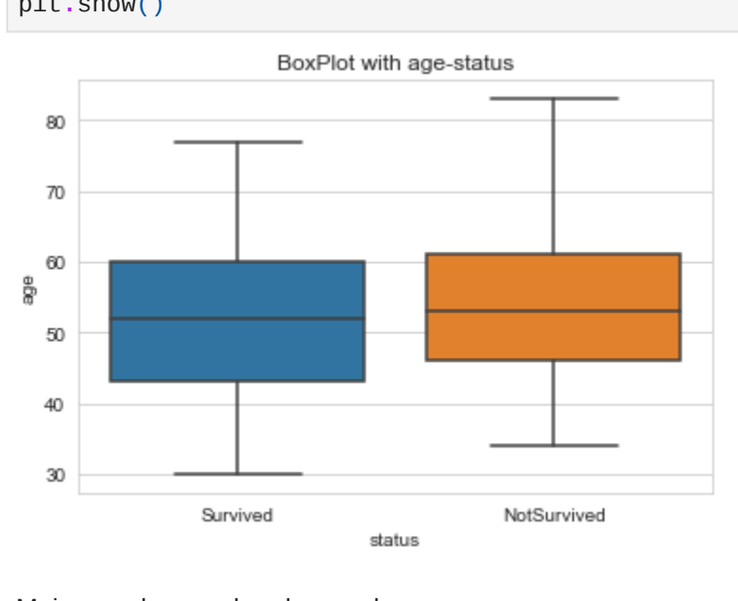
Observation
Observing the difference,It has gotten clearer that The petionts **chances of Survival decline with higher number of infected nodes**

BOX PLOT

In [58]:
```
sns.boxplot(x='status',y='year',data=df)
plt.title('Boxplot of status-year')
plt.grid
plt.show()
```

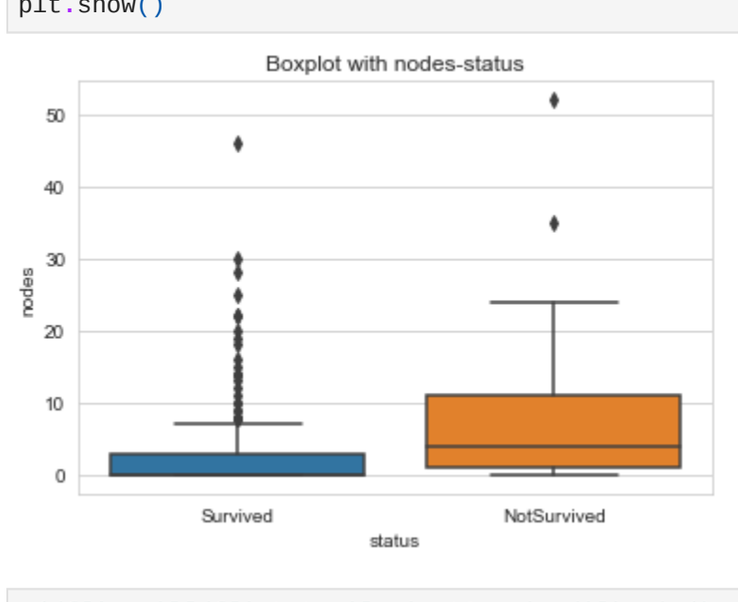


In [60]:
```
sns.boxplot(x='status',y='age',data=df)
plt.title('BoxPlot of age-status')
plt.grid
plt.show()
```



Major overlap can be observed.

In [61]:
```
sns.boxplot(x='status',y='nodes',data=df)
plt.title('Boxplot of nodes-status')
plt.show()
```



In [12]:
```
#(df['age'][df['status']=='NotSurvived']).min()
```

For Survived,

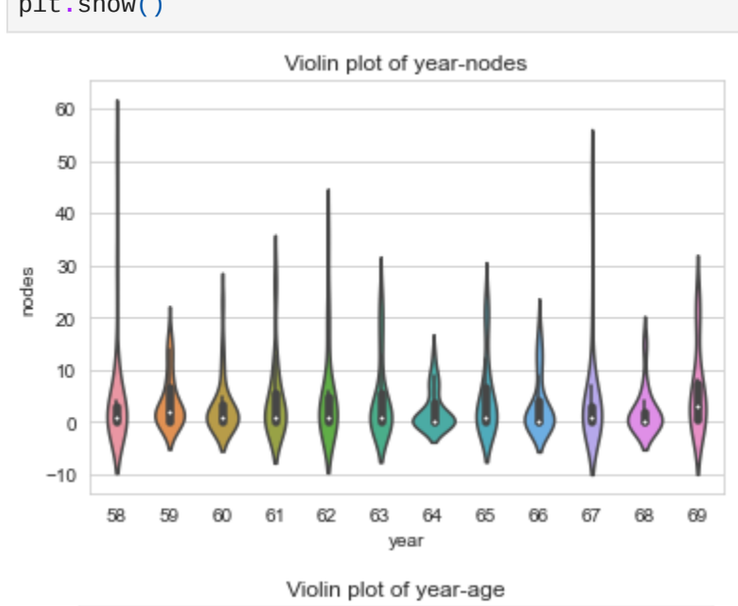
- The box plot with survived has values below 50th percentile all lying on 0,and therefore most of the petionts survived have 0 infected nodes.
- 75th percentile of petionts survived have about 3 or 4 infected nodes.
- The points beyond whiskers are less likely to be outliers because there are more than normal number of them.But the point at 46(i.e,Petiont with 46 infected nodes) is an extreme outlier,Because person generally has about 20-30 axillary nodes,person having 46 nodes is not possible.

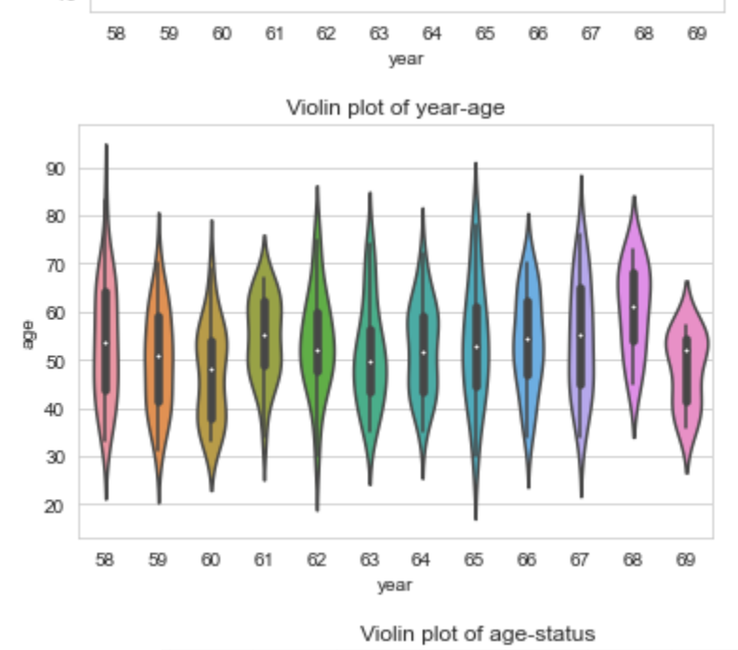
For NotSurvived,

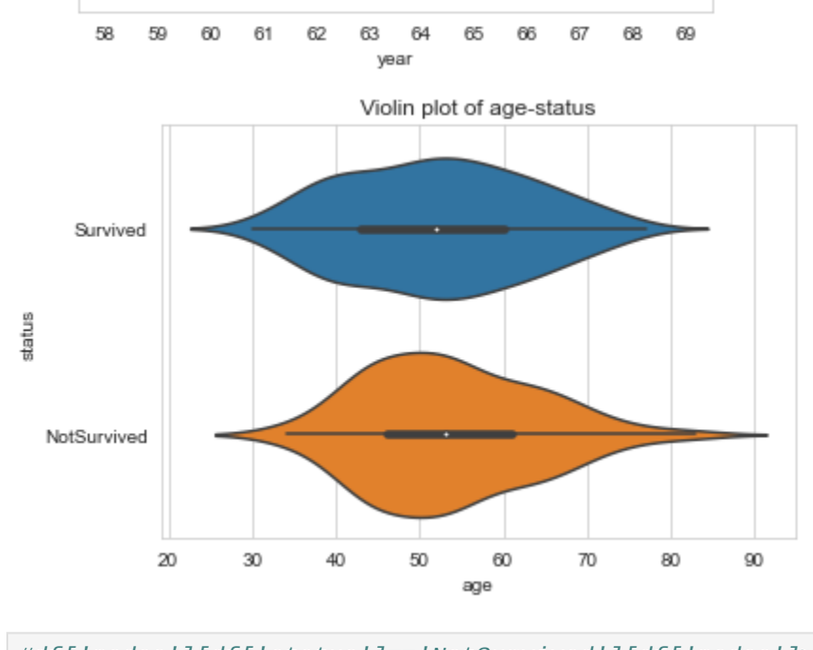
- 25th percentile of petionts not survived had about 0 or 1 infected nodes.
- About 50th percentile of petionts had infected nodes lesser than or equal to 4(<=: because the median lies on 4.5) and 75th percentile of petionts have lesser than equal to 11 infected nodes.
- The maximum number of infected nodes in petiont are 24.(excluding the extreme outliers).

VIOLIN PLOTS

In [62]:
```
sns.violinplot(data=df,x='year',y='nodes',size=30)
plt.title('Violin plot of year-nodes')
plt.show()

sns.violinplot(data=df,x='year',y='age',size=30)
plt.title('Violin plot of year-age')
plt.show()

sns.violinplot(data=df,x='age',y='status',size=30)
plt.title('Violin plot of age-status')
plt.show()
```







In [14]:
```
#df['nodes'][df['status']=='NotSurvived'][df['nodes']>=24]
```

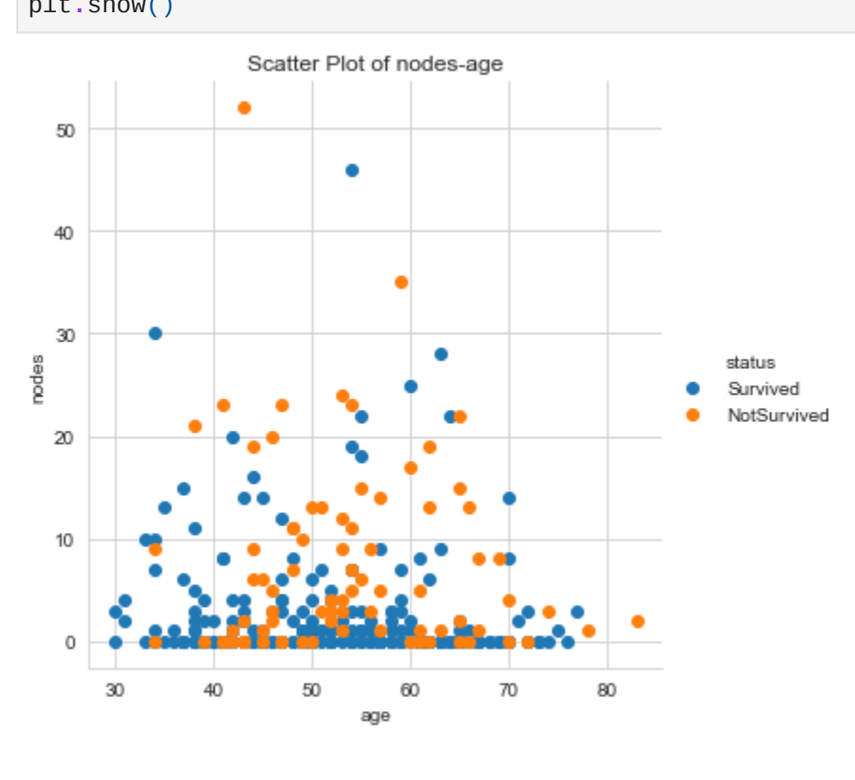
In [42]:
```
# We are undable to make sense of data because major overlap can be observed.
#So, We will try drawing conclusions through overall plot.
```

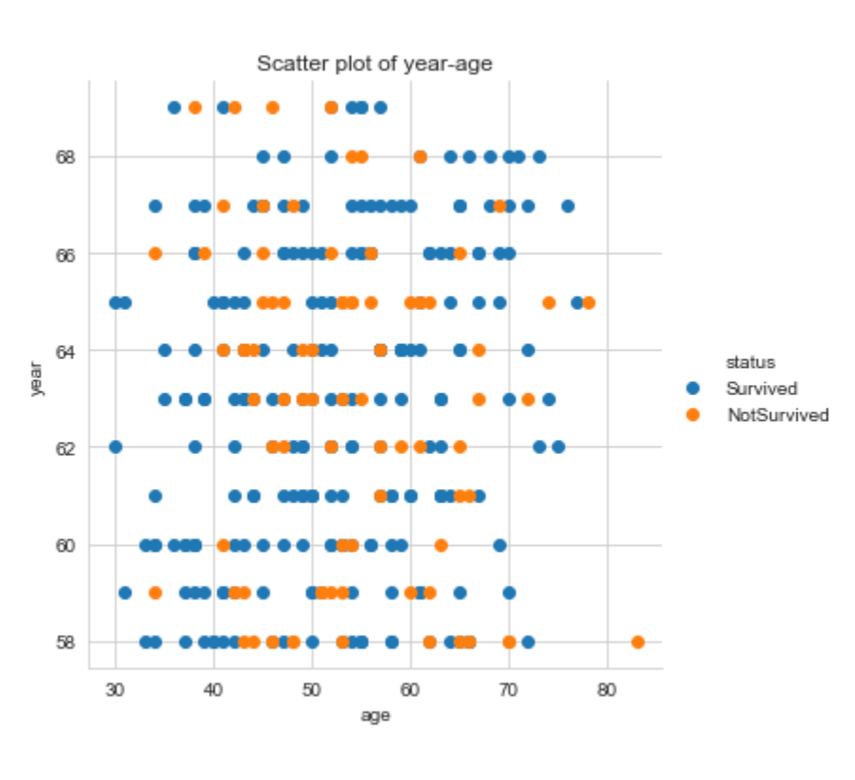
On plot of nodes-year:
We can observe the densities of plots are desributed mostly on points lesser than 11 and even more on 0 and 1,
It shows most of the petionts have lesser than 10 infected nodes and many of them have about 0 or 1 infected nodes.
On plot of age-year: We can observe that the densities of plots are destributed mostly on points from 45 to 60,
It shows most of petionts have age around 45 to 60.
On plot of age-status: We can observe that around 50 percent of people were about 43 to 62 year old.

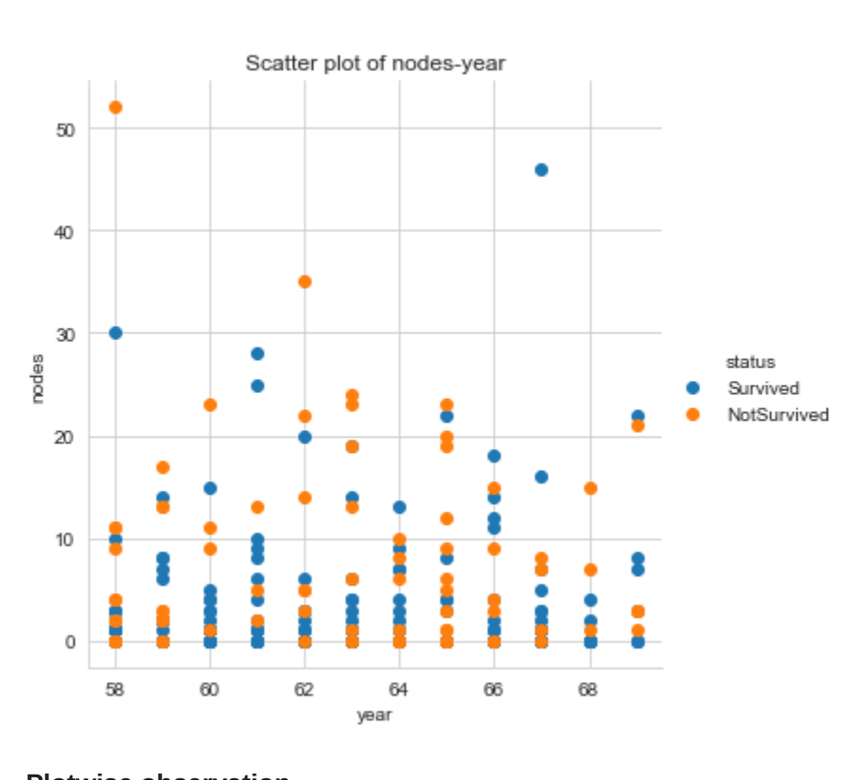
Bi-Variete Analysis

SCATTER PLOT

In [63]:
```
sns.set_style('whitegrid')

sns.FacetGrid(df,hue='status',size=5).map(plt.scatter,'age','nodes').add_legend().set(title='Scatter Plot of nodes-age')
plt.show()
print('\n')
sns.FacetGrid(df,hue='status',size=5).map(plt.scatter,'age','year').add_legend().set(title='Scatter plot of year-age')
plt.show()
print('\n')
sns.FacetGrid(df,hue='status',size=5).map(plt.scatter,'year','nodes').add_legend().set(title='Scatter plot of nodes-year')
plt.show()
```







Plotwise observation
Plot of nodes-age and year-age:

- Irrespective of nodes and year people having age lesser than 40 have higher chances of survival.

Plot of nodes-age and nodes-year:

- Many of petionts have lesser than 10 infected nodes & Most of them have 0 infected nodes.

We can observe that for observation of **age** , the **age-year** plot shows clearer representation.
Whereas,for observation of **number of nodes** the **nodes-age** plot shows better representation.

PAIR PLOT

In [16]:
```
#Pairplot

sns.set_style("whitegrid");
sns.pairplot(df,hue='status',size=5);
plt.show()
```

The scatter plot with **age** and **nodes** shows:
1.lesser the cancer spread across nodes, **higher the chances of survival.**
2.Plot of nodes-age shows better representation.

... ...

Overall Conclusion through Analysis:

- Through Description:
  1. The Youngest Petiont is of 30 years old,whereas Eldest is 83 years old.
  2. The survival rate of petionts is about 73%.
- Through Uni-Variate & Bi-Variate Analysis:
  1. petiont has higher chances of survival if petiont is lesser than 40 year old or petiont has lesser age.
  2. petiont has lesser infected nodes,even better if petiont has lesser than 4 infected nodes.
  3. Most of the surviving petionts have 0 infected nodes.
  4. Most of the petionts (irrespective of status) have lesser than 11 infected nodes.
  5. Most of petionts(around 50%) are 43 to 63 years old

In [17]:
```
%%html
<style>
table {float:left}
</style>
```