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

import warnings
warnings.filterwarnings("ignore")
```

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

The dataset contains cases from a study that was conducted between 1958 and 1970 at the Univerity of Chicago's Billings Hospital on the survival of the patients who had under gone breast cancer surgery

we have 3 feature and 1 class label

1. **Age**- age of the patient at the time of operation
2. **Year** - Year in which the patient had operationin 90's . If 64 is written then in 1964.
3. **Nodes**- Number of positive Auxillary Nodes detected. This are knowns as "Auxillary lymph Nodes" whose function is to filter fluids before they are eventually released into the bloodstream . Having cancer cells in the nodes tells us that cancer might have spread to other body parts also.
4. **Survival status** - This is the target varible having two values 1 and 2.

```
1 indicates the patient survived 5yrs or longer Post operation
2 indicates the patient died with in 5yrs
```

The datasets have NO columns name specified hence we name each of the columns

In [3]:
```
df.columns=['Age','Year','Nodes','Survival']
```

In [4]:
```
df.head()
```

Out[4]:

	Age	Year	Nodes	Survival
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

AGENDA : Given a patient with some Age, Year of Opperation, Numbr of Lymph Nodes Detected, we have to predict or

conclude the chances wheather the patient will live less than 5 yrs or more than 5 yrs

In [5]:
```
df.shape
```

Out[5]:
```
(305, 4)
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
Age         0
Year        0
Nodes       0
Survival    0
dtype: int64
```

In [7]:
```
df.duplicated().value_counts()
```

Out[7]:
```
False    288
True      17
Name: count, dtype: int64
```

No need for removal of duplicates since more than two people can encounter same situations

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 305 entries, 0 to 304
Data columns (total 4 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Age       305 non-null    int64
 1   Year      305 non-null    int64
 2   Nodes     305 non-null    int64
 3   Survival  305 non-null    int64
dtypes: int64(4)
memory usage: 9.7 KB
```

In [9]:
```
df.Survival.value_counts()
```

Out[9]:
```
Survival
1    224
2     81
Name: count, dtype: int64
```

In [10]:
```
df=df.replace([1,2],[0,1])
```

In [11]:
```
df.Survival.value_counts()
```

Out[11]:
```
Survival
0    224
1     81
Name: count, dtype: int64
```

In [12]:
```
df.describe()
```

Out[12]:

	Age	Year	Nodes	Survival
count	305.000000	305.000000	305.000000	305.000000
mean	52.531148	62.849180	3.839344	0.265574
std	10.744024	3.254078	7.283978	0.442364
min	30.000000	58.000000	0.000000	0.000000
25%	44.000000	60.000000	0.000000	0.000000
50%	52.000000	63.000000	0.000000	0.000000
75%	61.000000	66.000000	4.000000	1.000000
max	83.000000	69.000000	52.000000	1.000000

In [13]:
```
#MIN(AGE)=30, MAX(AGE)=83, YEAR =[1958-1969], NODES=[0,52] ,SUVIVAL=[0,1]

# NODES HAVE HIGHER COUNT DENSITY BETWEEN 0 AND 4(75% PERCENTILE)
```

UNIVARIATE ANALYSIS

1.SURVIAVAL

In [14]:
```
df.Survival.value_counts()
```

Out[14]:
```
Survival
0    224
1     81
Name: count, dtype: int64
```

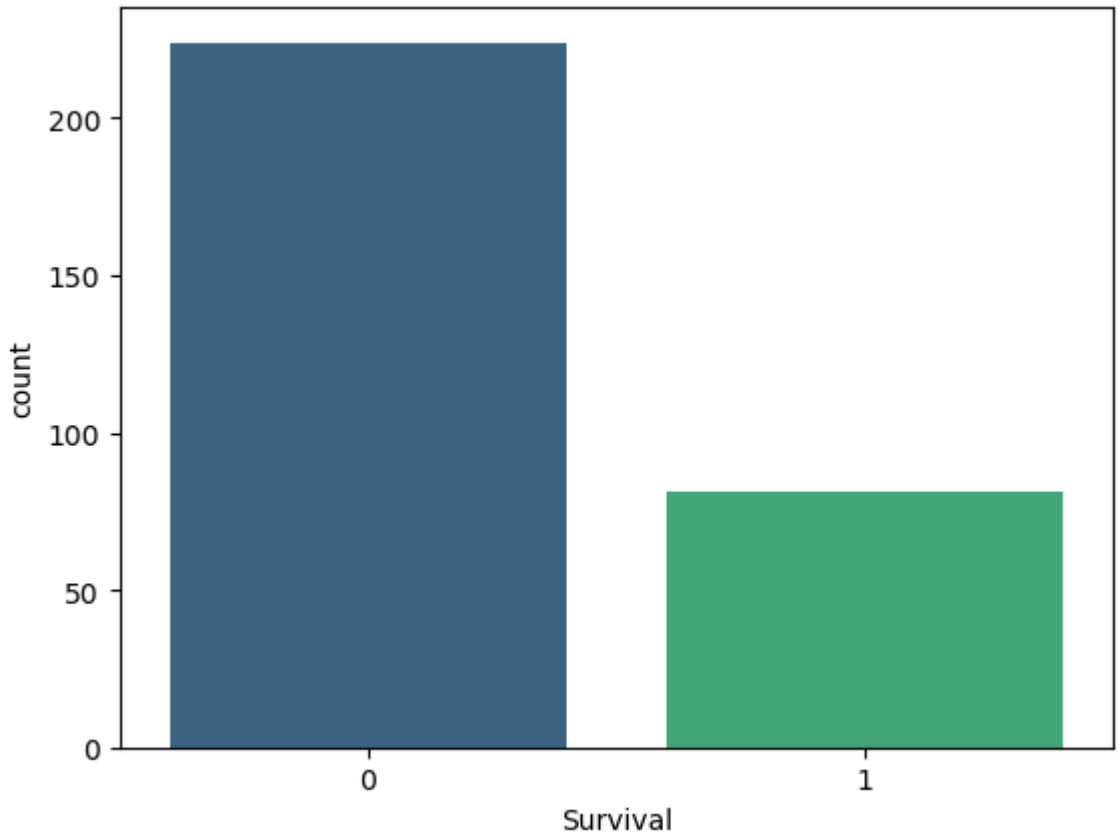
In [15]:
```
#converting percentage
df.Survival.value_counts(1)
```

Out[15]:
```
Survival
0    0.734426
1    0.265574
Name: proportion, dtype: float64
```

only 27% of people who will bw living for less than 5yrs, which means that the dataset is imbalance

In [16]:
```
sns.countplot(x='Survival',data=df,palette='viridis')
```

Out[16]:
```
<Axes: xlabel='Survival', ylabel='count'>
```



AGE

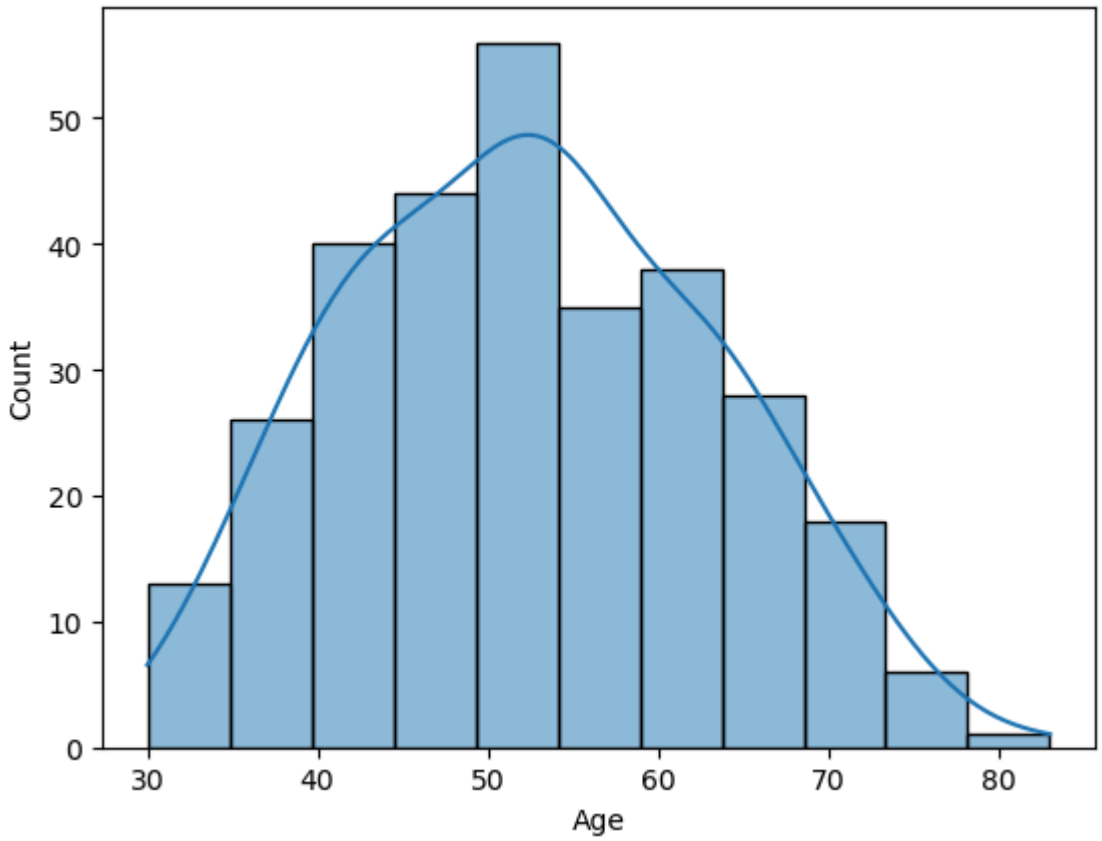
In [17]:
```
df.Age.describe()
```

Out[17]:
```
count    305.000000
mean      52.531148
std       10.744024
min       30.000000
25%       44.000000
50%       52.000000
75%       61.000000
max       83.000000
Name: Age, dtype: float64
```

In [18]:
```
print(df.Age.skew())
sns.histplot(x='Age', data=df,binwidth=5,kde=True)
```

```
0.15898611605406873
```

Out[18]:
```
<Axes: xlabel='Age', ylabel='Count'>
```

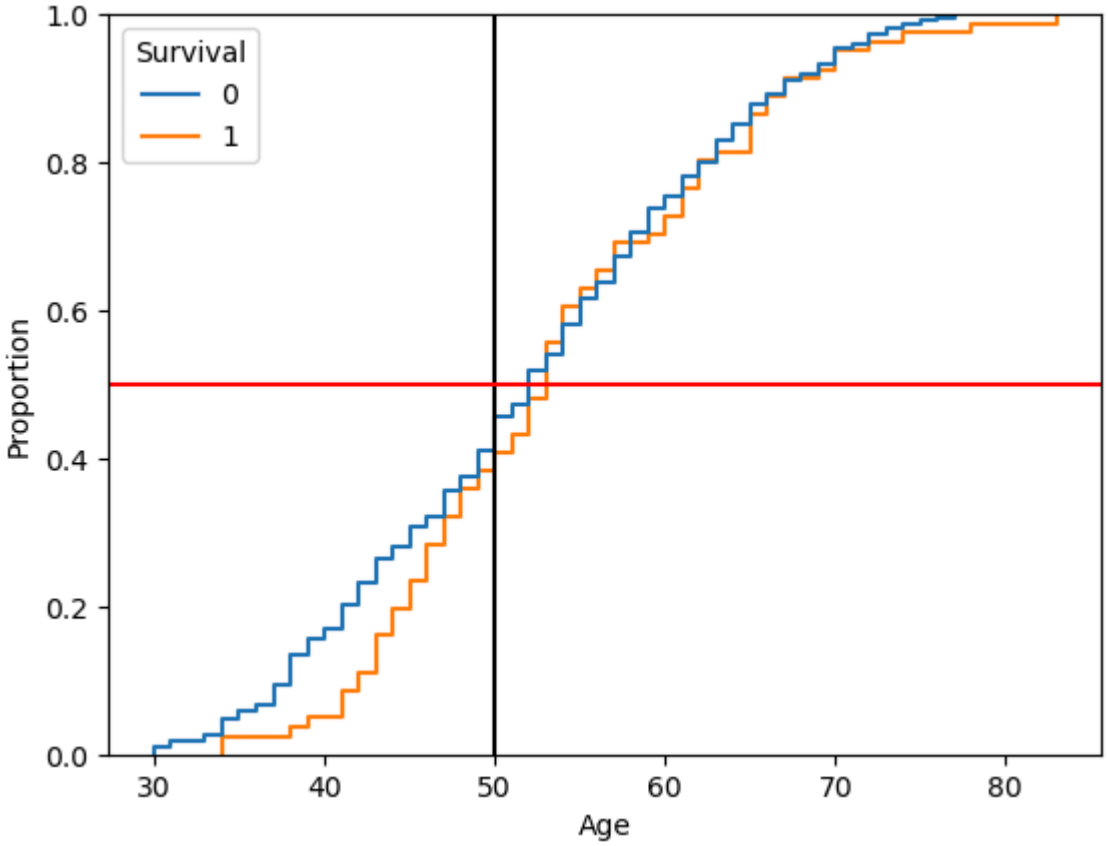


kde captures even the smal and minor details of a dataset whereas hitogram might miss those details like for example a datset has BIMODEL this fact will be captured by KDE, also histogram groups the data into bins and KDE analyzes the behaviour of this datapoints in each of this bins, the bin which has more no. of data points will have higher density and vice-versa.

From the above diagram we can conclude that the maximum number of people who were operated lie with in the age group of 50-55

In [19]:
```
sns.ecdfplot(x='Age',data=df,hue='Survival')
plt.axvline(50,c='black')
plt.axhline(0.5,c='red')
```

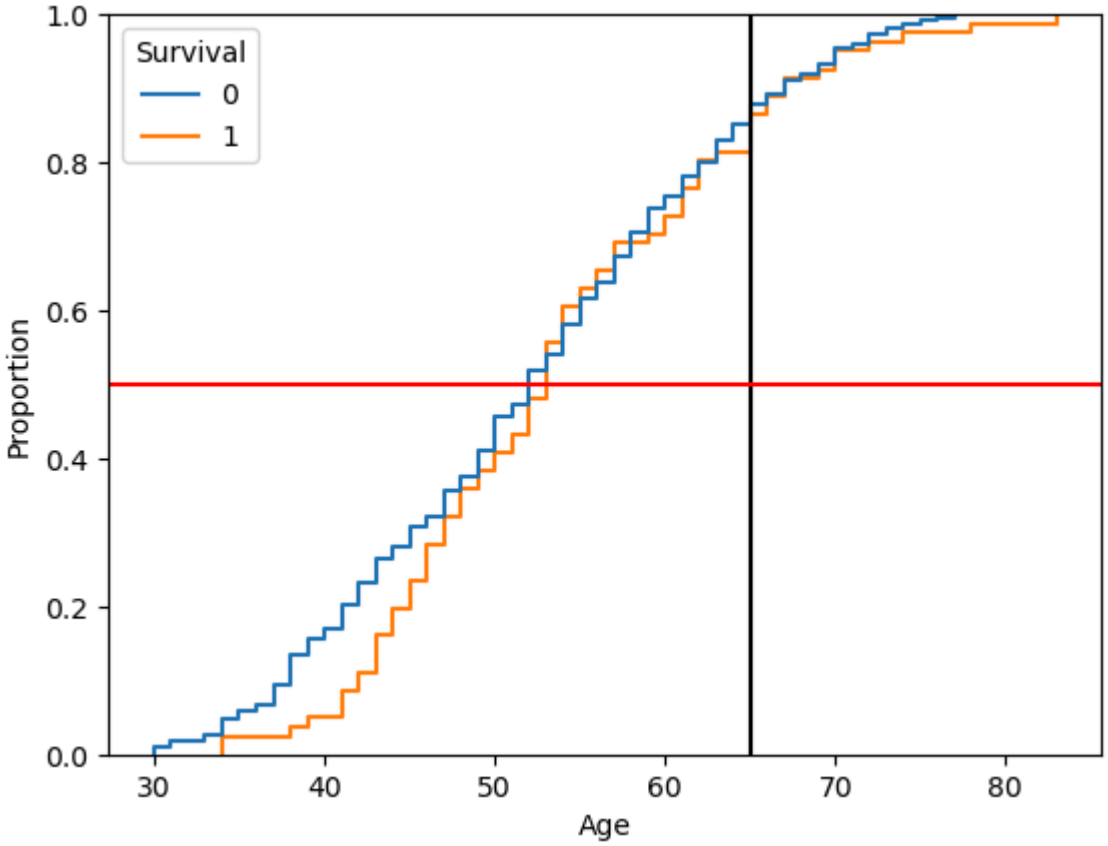
Out[19]:
```
<matplotlib.lines.Line2D at 0x19b944c1fd0>
```



Here you can get answer to age group with 50% chance of survival? and age group with 50% cahnce of non survival?

In [20]:
```
sns.ecdfplot(x='Age',data=df,hue='Survival')
plt.axvline(65,c='black')
plt.axhline(0.5,c='red')
```

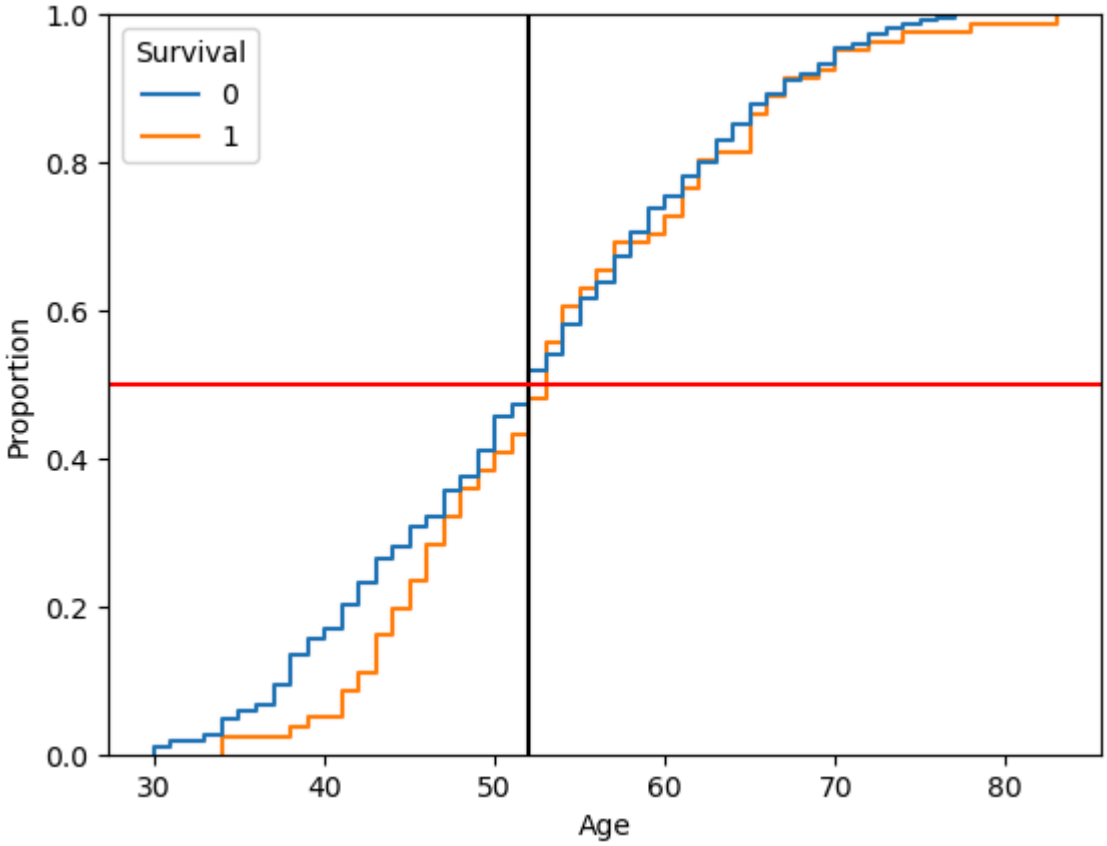
Out[20]:
```
<matplotlib.lines.Line2D at 0x19b955d7b00>
```



80% of the people were 65yrs or beleow

In [21]:
```
sns.ecdfplot(x='Age',data=df,hue='Survival')
plt.axvline(52,c='black')
plt.axhline(0.5,c='red')
```

Out[21]:
```
<matplotlib.lines.Line2D at 0x19b95696cf0>
```

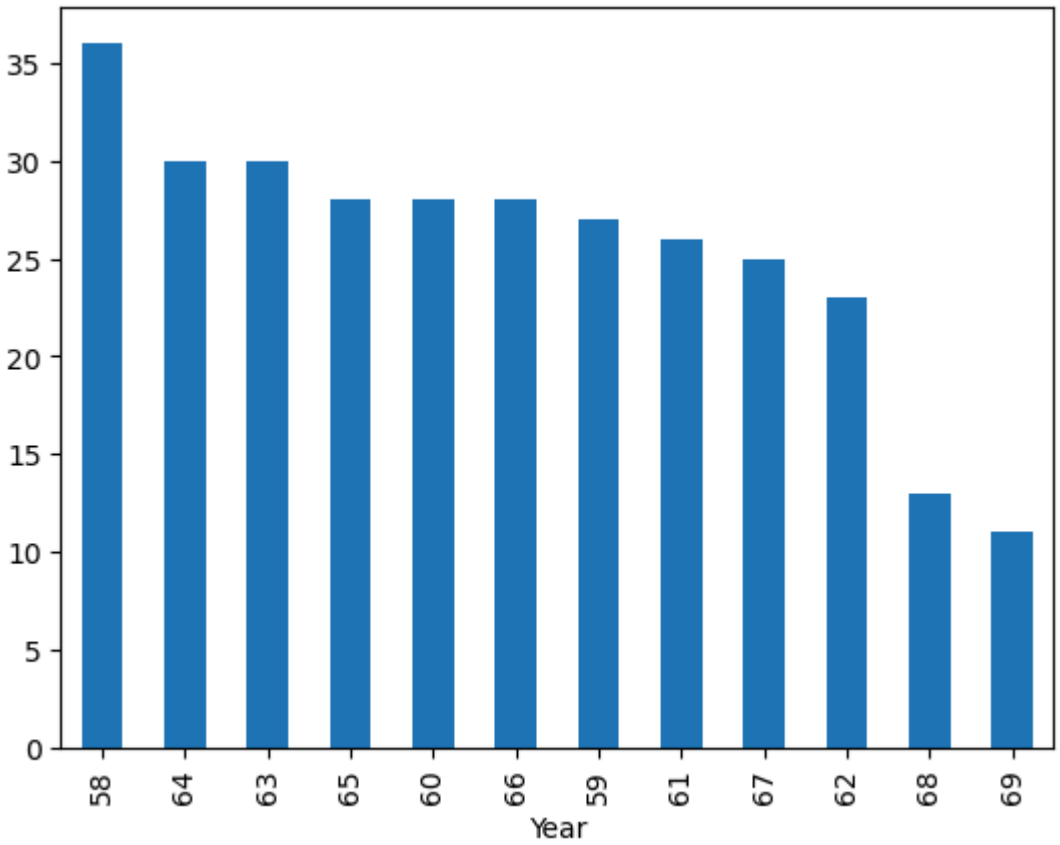


50% of the people were 52 yrs or beleow

OPERATION YEAR

In [22]:
```
df.Year.value_counts().plot(kind='bar')
```

Out[22]:
```
<Axes: xlabel='Year'>
```



from here we can get an idea why no. of doctors chnages year wise ? why we see a drastic fall in no. of doctors?

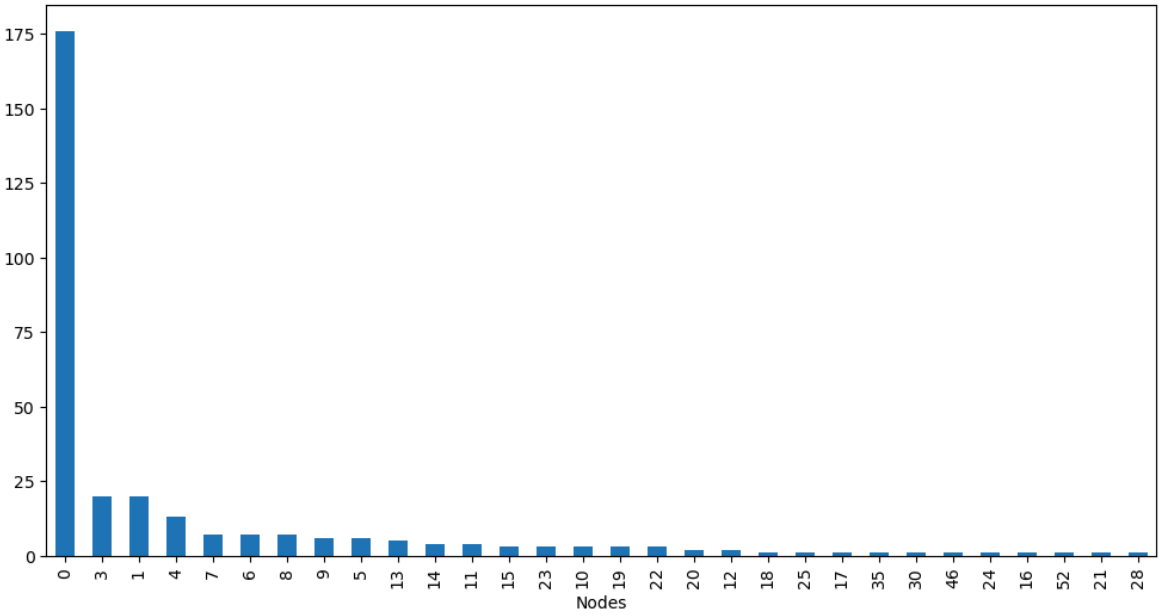
LYMPH NODE

In [23]:
```
df.Nodes.describe()
```

Out[23]:
```
count    305.000000
mean       3.839344
std        7.283978
min        0.000000
25%        0.000000
50%        0.000000
75%        4.000000
max       52.000000
Name: Nodes, dtype: float64
```

In [24]:
```
plt.figure(figsize=(12,6))
df.Nodes.value_counts().plot(kind='bar')
```

Out[24]:
```
<Axes: xlabel='Nodes'>
```

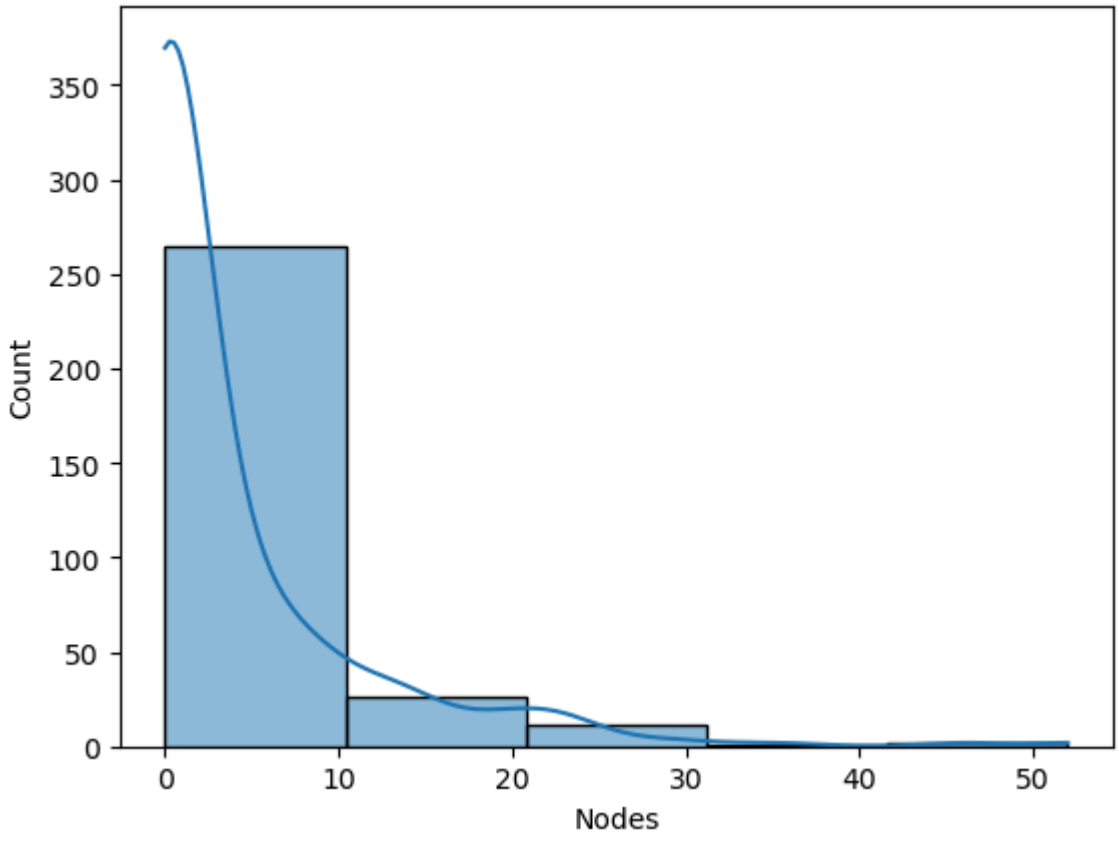


From here we observe that maximum people has decreasing lymph nodes, also we observe a decreasing trend

In [26]:
```
print(df.Nodes.skew())
sns.histplot(x='Nodes', data=df,binwidth=10,kde=True)
```

```
2.940405369162834
```

Out[26]: <Axes: xlabel='Nodes', ylabel='Count'>

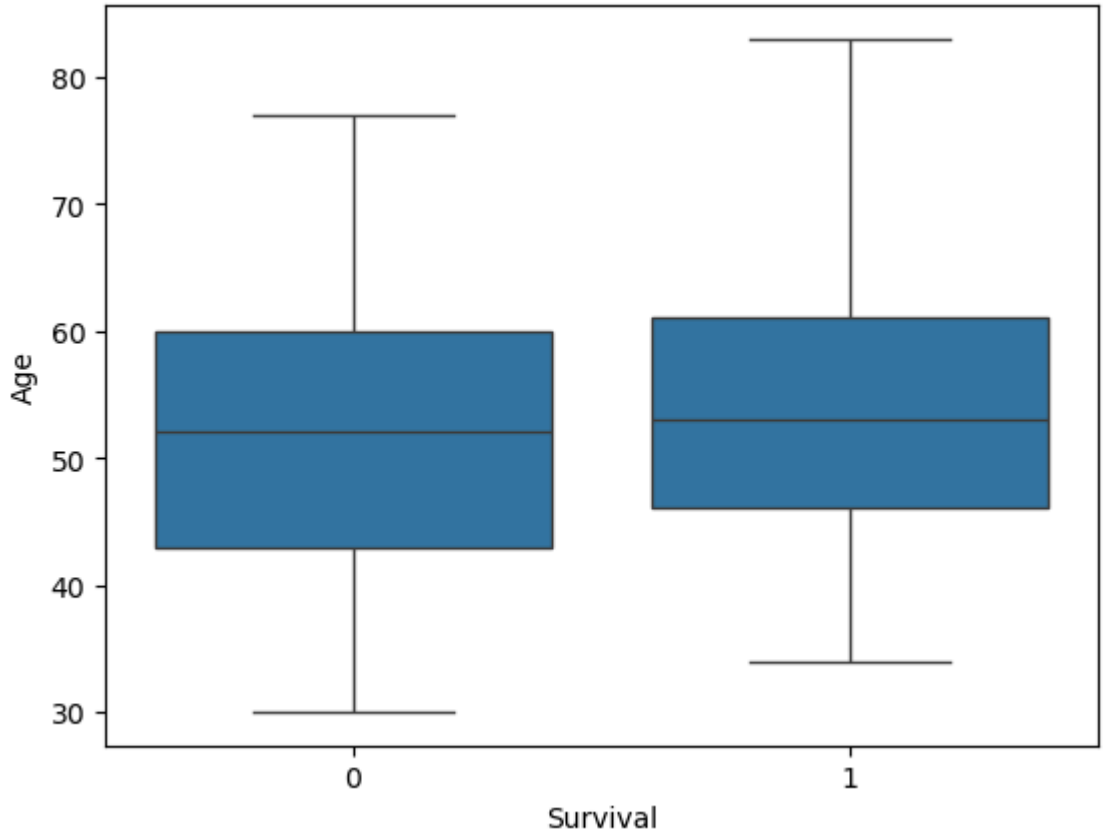


we observe here that the distribution is rightly skewed.. also, Maximum Density count for Nodes is between 0-10

NOW WE DO BIVARIATE ANALYSIS

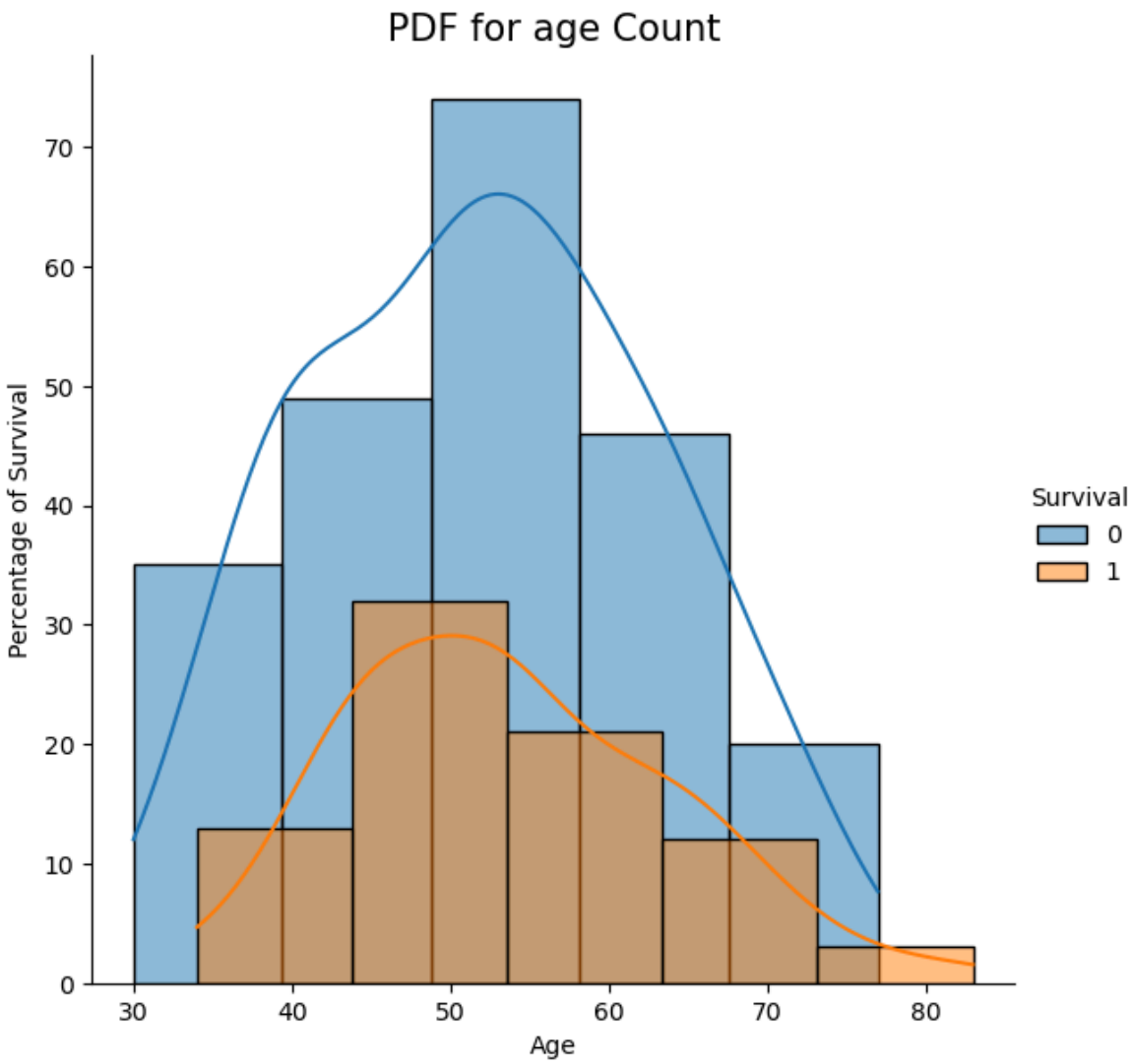
In [27]:
```
# Age-Survival
sns.boxplot(x='Survival',y='Age',data=df)
```

Out[27]: <Axes: xlabel='Survival', ylabel='Age'>



- *here 1st we observe e have no outliers*
- *people with age>77will actually live lesser than 5 yrs*
- *people with age lesser than age<35 will actually will more than 5yrs*

In [31]:
```
sns.FacetGrid(df,hue='Survival',height=6).map(sns.histplot,'Age',bins=5,kde=True
plt.xlabel('Age')
plt.ylabel('Percentage of Survival')
plt.title('PDF for age Count', size=15)
plt.show()
```



- *maximum % of people living lesser than 5yrs are between 45-55*
- *people between 30-33yrs old are living more than 5 yrs*
- *people with age>77will actually live lesser than 5 yrs*

In [35]:
```
df.groupby('Age')['Survival'].sum().sort_values(ascending=False)
```

Out[35]:
```
Age
53    6
46    4
43    4
65    4
54    4
52    4
47    3
61    3
45    3
44    3
48    3
41    3
57    3
62    3
50    2
49    2
56    2
51    2
55    2
42    2
66    2
67    2
70    2
34    2
60    2
69    1
72    1
74    1
63    1
78    1
83    1
59    1
39    1
38    1
36    0
33    0
77    0
76    0
75    0
35    0
73    0
71    0
58    0
37    0
68    0
40    0
31    0
64    0
30    0
Name: Survival, dtype: int64
```

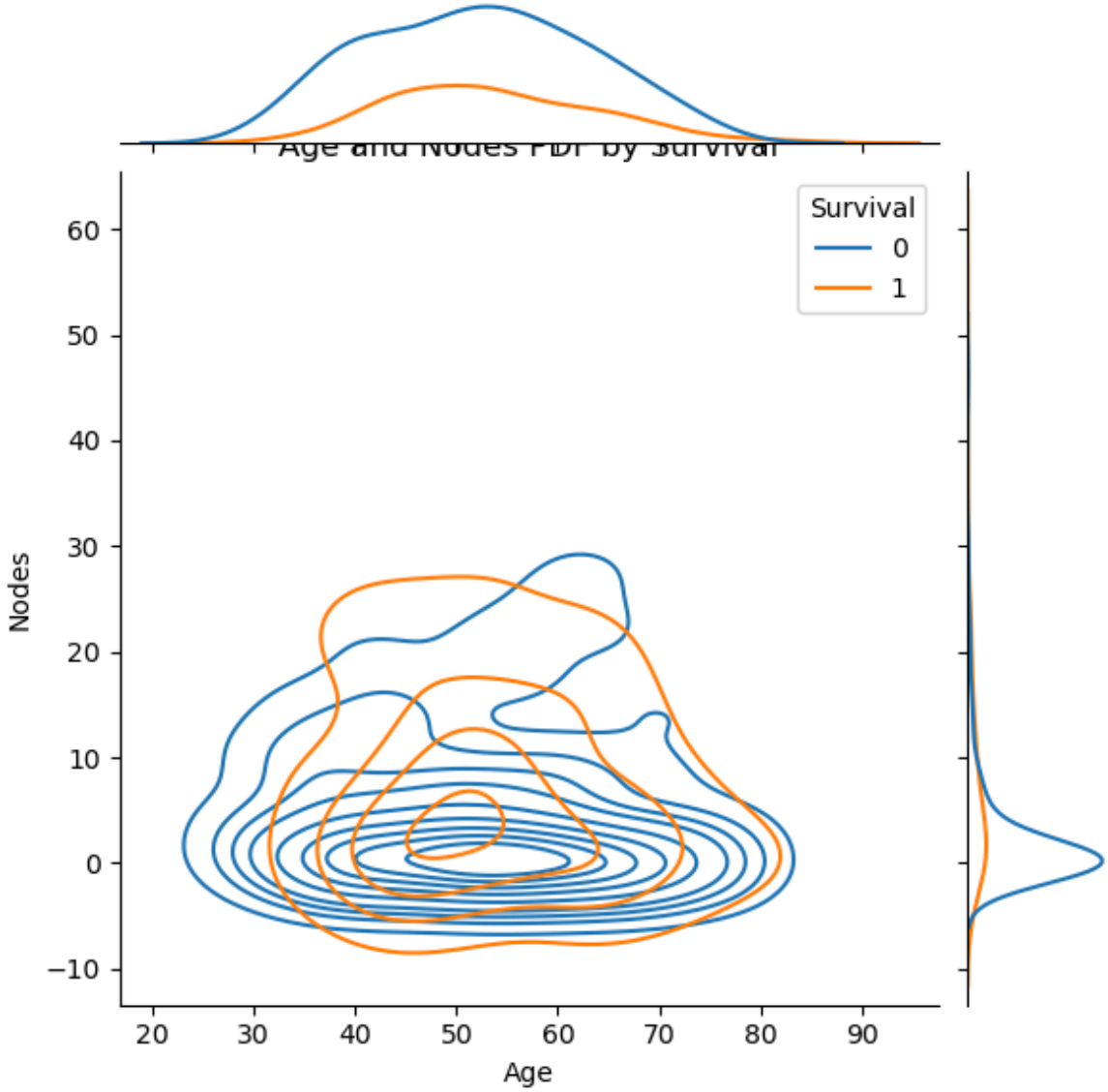
- *6 people will live lesser than 5yrs who are in the afe group of 53*

In [36]:
```
df.Age.value_counts().sort_values(ascending=False)
```

Out[36]:
```
Age
52    14
54    13
50    12
57    11
53    11
47    11
43    11
65    10
38    10
49    10
41    10
55    10
42     9
45     9
61     9
63     8
59     8
70     7
46     7
48     7
44     7
34     7
62     7
56     7
58     7
37     6
51     6
60     6
39     6
67     6
66     5
64     5
72     4
69     4
40     3
31     2
33     2
35     2
36     2
73     2
68     2
30     2
74     2
71     1
75     1
76     1
77     1
78     1
83     1
Name: count, dtype: int64
```

In [40]:
```
#Age -Nodes-Survival
plt.figure(figsize=(12,6))
sns.jointplot(x='Age',y='Nodes',data=df,hue='Survival',kind='kde')
plt.title('Age and Nodes PDF by Survival')
plt.show()
```

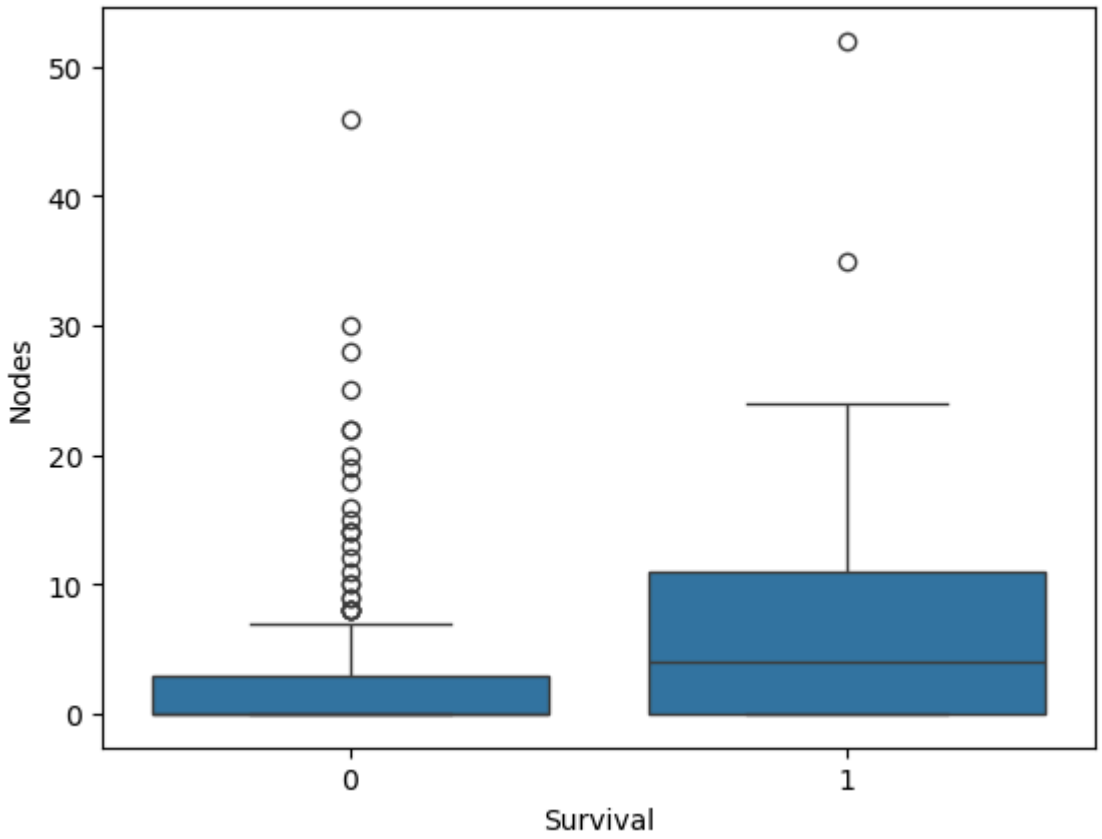
```
<Figure size 1200x600 with 0 Axes>
```



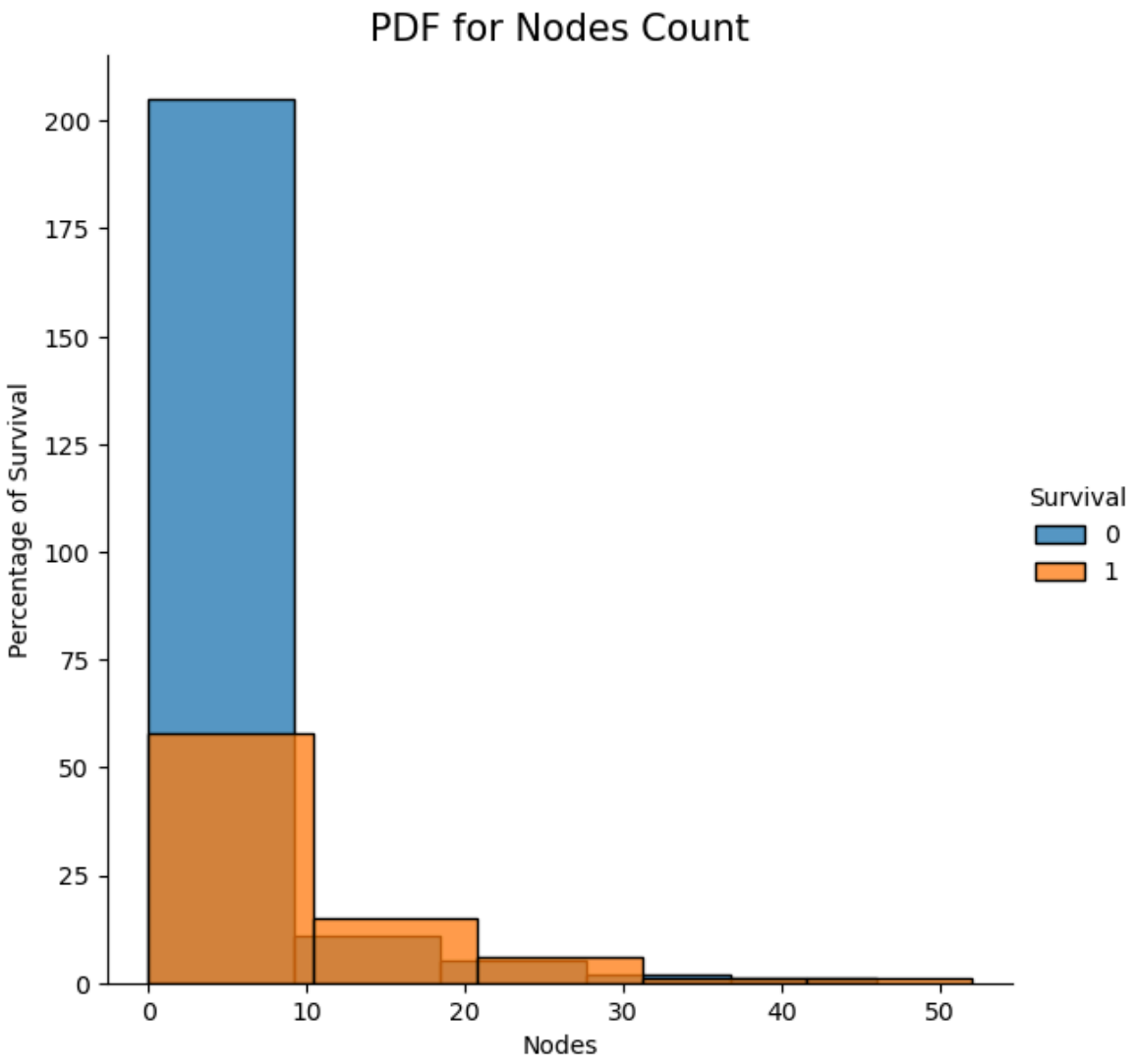
AGE[43,53]and NODES[10-52] will live lesser than 5yrs

In [41]:
```
sns.boxplot(x='Survival',y='Nodes',data=df)
```

Out[41]:
```
<Axes: xlabel='Survival', ylabel='Nodes'>
```



In [44]:
```
sns.FacetGrid(df,hue='Survival',height=6).map(sns.histplot,'Nodes',bins=5).add_l
plt.xlabel('Nodes')
plt.ylabel('Percentage of Survival')
plt.title('PDF for Nodes Count', size=15)
plt.show()
```



In [45]:
```
df.groupby('Nodes')['Survival'].sum().sort_values(ascending=False)
```

Out[45]:
```
Nodes
0     27
3      7
1      5
13     4
5      4
9      4
4      3
6      3
23     3
11     3
15     2
19     2
8      2
7      2
24     1
20     1
22     1
21     1
35     1
52     1
17     1
14     1
12     1
10     1
18     0
25     0
28     0
30     0
46     0
16     0
Name: Survival, dtype: int64
```

In [46]:
```
df.Nodes.value_counts().sort_values(ascending=False)
```

Out[46]:

Nodes	
0	176
1	20
3	20
4	13
7	7
6	7
8	7
9	6
5	6
13	5
14	4
11	4
10	3
22	3
19	3
23	3
15	3
20	2
12	2
18	1
25	1
17	1
35	1
30	1
46	1
24	1
16	1
52	1
21	1
28	1

Name: count, dtype: int64

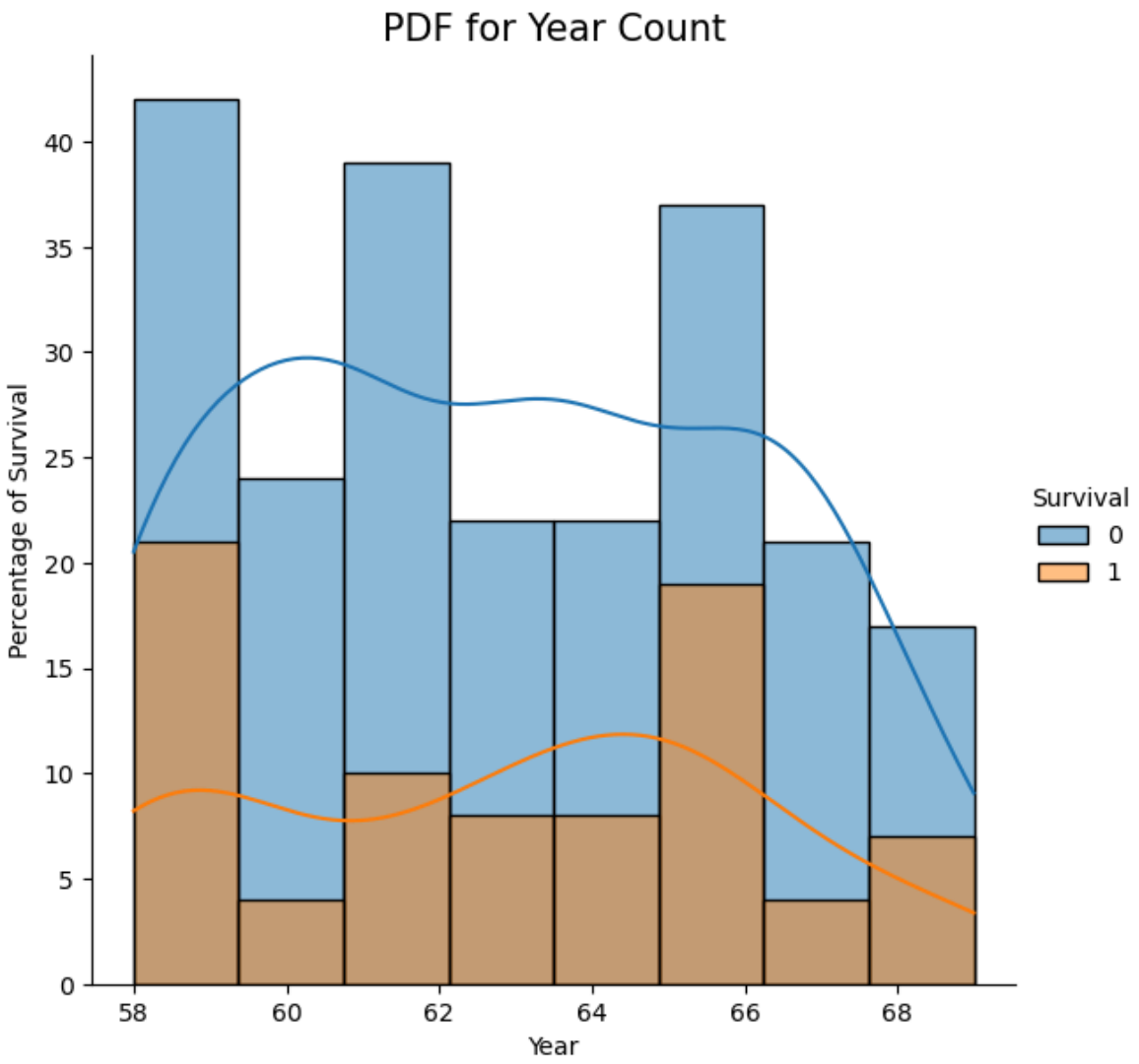
27/176 * 100 = 15%(approx) 7/20*100 = 35%(approx) 3/13*100 = 23%(approx) 4/6*100 = 67%(approx)

- *people with Nodes 0 has 15% chance of dieing before 5yrs*
- *people with Nodes 1 will have 25% of chance of surving before 5yrs*
- *people with Nodes 3 will have 35% of chance of Surving before 5yrsor lets say 35% chance of surving lesser than 5yrs*
- *people with Nodes 4 will have 23% of chance of Surving before 5yrsor lets say 23% chance of surving lesser than 5yrs*
- *people with Nodes 5 will have 67% of chance of Surving before 5yrsor lets say 67% chance of surving lesser than 5yrs*

and so on..

and from this we get that there is high chance that after **10** the people will not survive

In [50]:
```
sns.FacetGrid(df,hue='Survival',height=6).map(sns.histplot,'Year',bins=8,kde=Tru
plt.xlabel('Year')
plt.ylabel('Percentage of Survival')
plt.title('PDF for Year Count', size=15)
plt.show()
```



In [51]:
```
df.groupby('Survival')['Year'].value_counts().unstack()
```

Out[51]:

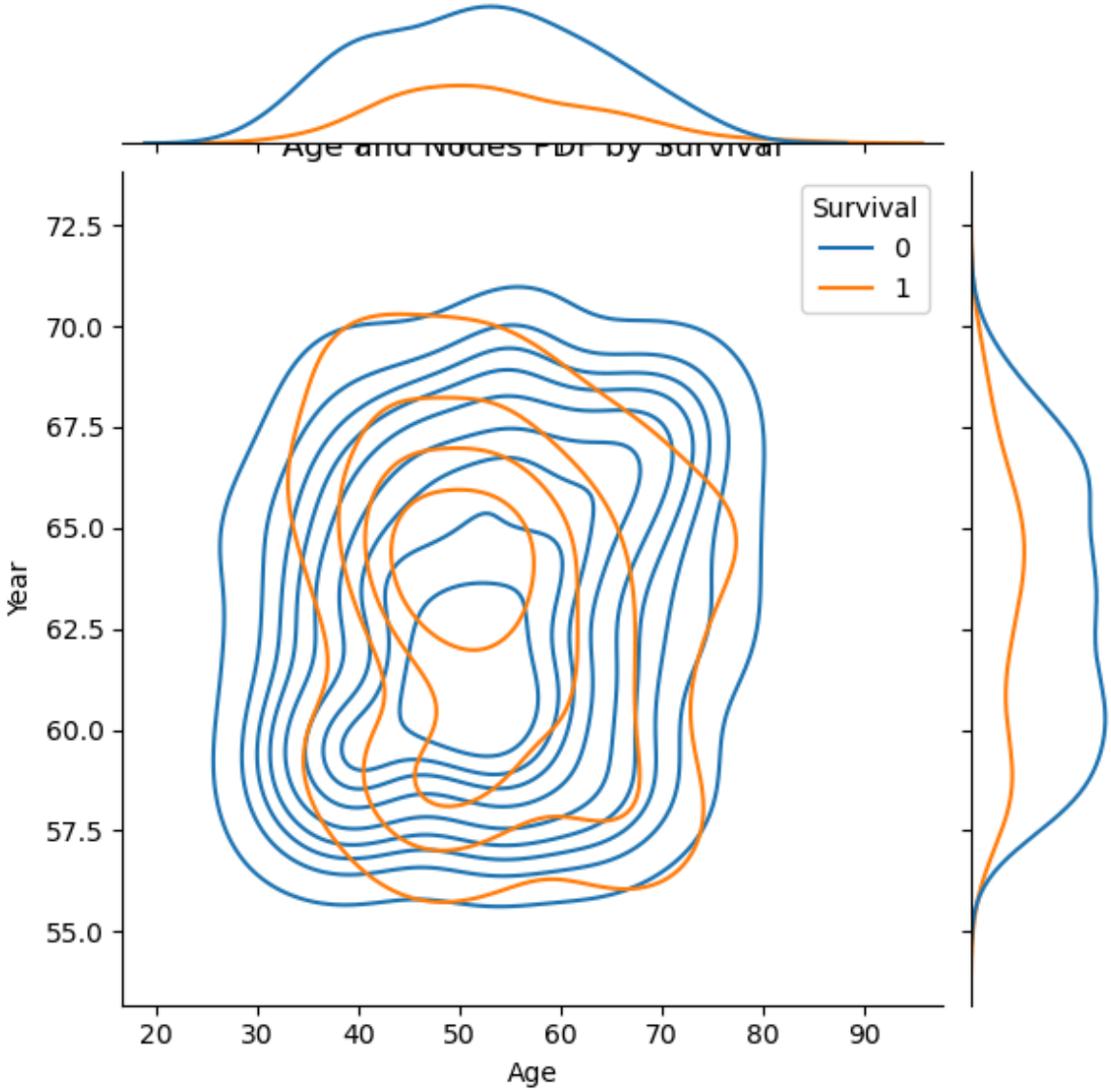
Survival	58	59	60	61	62	63	64	65	66	67	68	69
0	24	18	24	23	16	22	22	15	22	21	10	7
1	12	9	4	3	7	8	8	13	6	4	3	4

- *1965- we find 46% of the people died [13/28x100]*
- *1958- we find 33% of the people died*
- *1959- we find 45% of the people died*

33% of people died before 5yrs when operated between 1965-66

In [52]:
```
plt.figure(figsize=(12,6))
sns.jointplot(x='Age',y='Year',data=df,hue='Survival',kind='kde')
plt.title('Age and Nodes PDF by Survival')
plt.show()
```

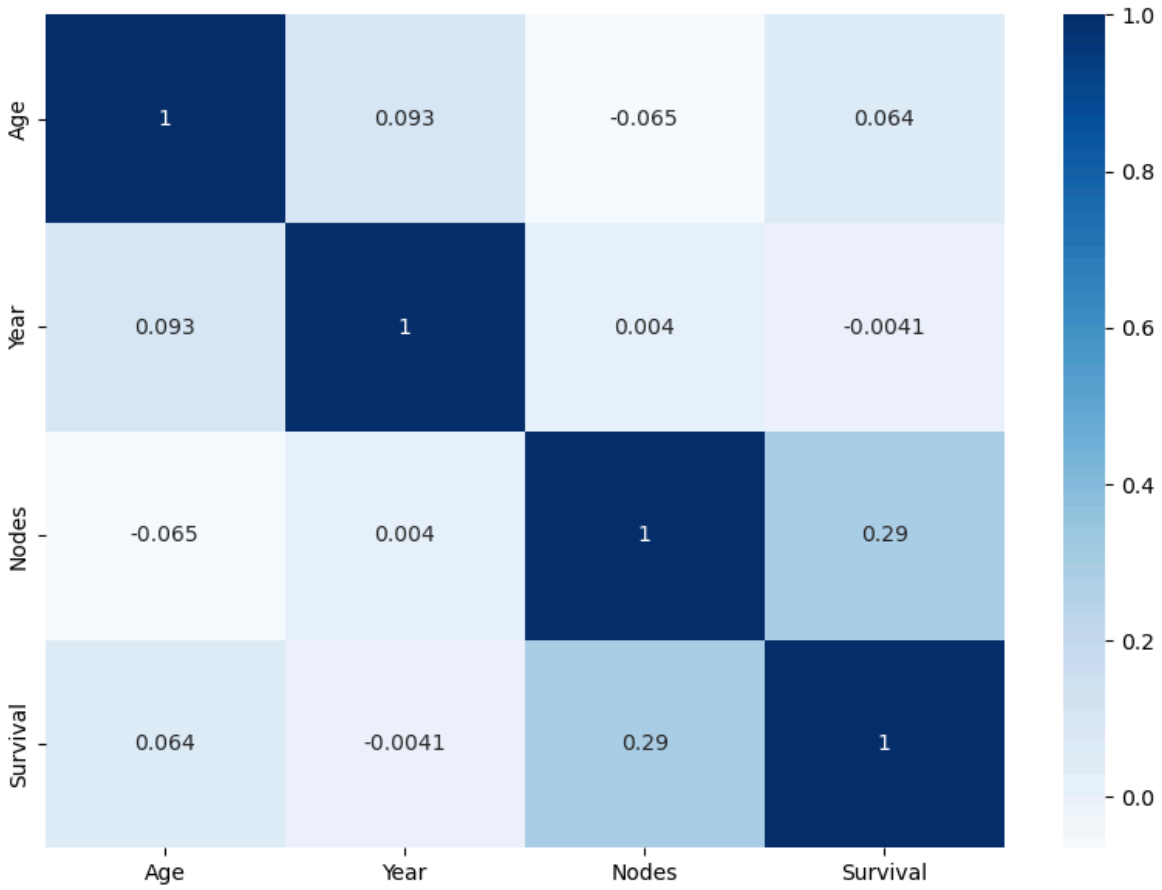
```
<Figure size 1200x600 with 0 Axes>
```



MULTIVARIATE ANALYSIS

In [55]:
```
plt.figure(figsize=(10,7))
corr=df.corr()
sns.heatmap(corr,xticklabels=corr.columns,yticklabels=corr.columns,cmap='Blues',
```

Out[55]: <Axes: >



nodes and survival are correlated by 29%

In [ ]: