# High level statistics of the dataset:

In [32]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_years"]
hbdata["Survived_after_5_years"] = hbdata["Survived_after_5_years"].map({1:"yes", 2:"no"})
print (hbdata.shape)#This would tell us number of rows and columns
print (hbdata["Survived_after_5_years"].value_counts())#This tells us totol number of cases we had
how many survived longer
#than five years and how many didnt not survived more than 5 years
```

```
(305, 4)
yes    224
no      81
Name: Survived_after_5_years, dtype: int64
```

# High level statistics of the dataset

- Total data points : 305 (total number of rows)
- Number of features : 3 (total columns - 1)
- Number of clases : 2 (total number of people survived less and more than 5 years)
- Total number of people who survived more than 5 years of 305 people : 224
- Total number of people who didn't survive more than 5 years of 305 people : 81
- Thus we have imbalanced set of data

# Lets Perform 2D scatter plot first

- Total possible scatter plots 1)Age n Yr of Oprtn 2)Age n +ve Nodes 3)Yr of Opertn n +ve Nodes

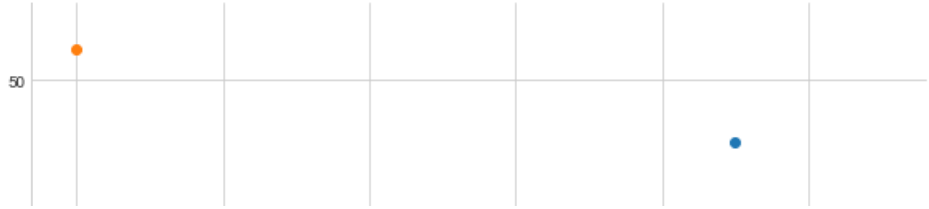
# Scatter Plot Year of Operation and Positive nodes (Plot 1)

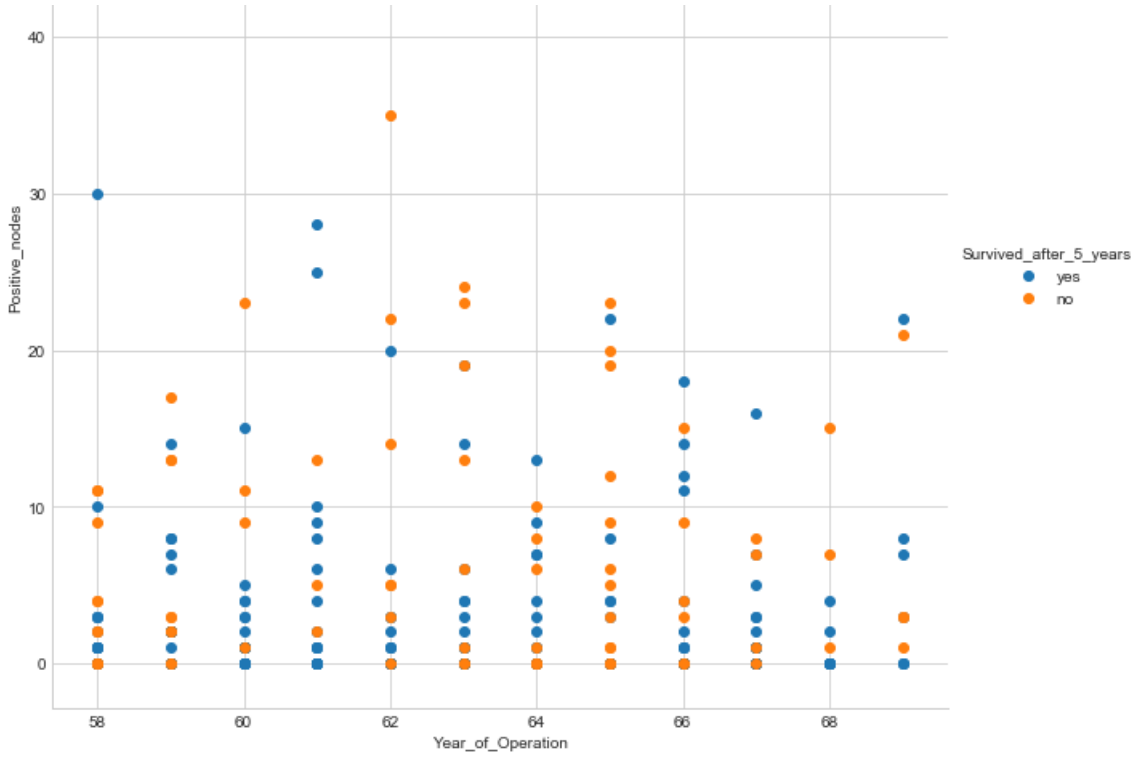
In [4]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_years"]
hbdata["Survived_after_5_years"] = hbdata["Survived_after_5_years"].map({1:"yes", 2:"no"})

sns.set_style("whitegrid");
sns.FacetGrid(hbdata, hue="Survived_after_5_years", size=8.5) \
   .map(plt.scatter, "Year_of_Operation", "Positive_nodes") \
   .add_legend();
plt.show();
```






# Observations:

- The data is quite scattered and personally i am finding quite difficult to come to any conclusion with help of this

# Scatter Plot Age and Positive nodes (Plot 2)

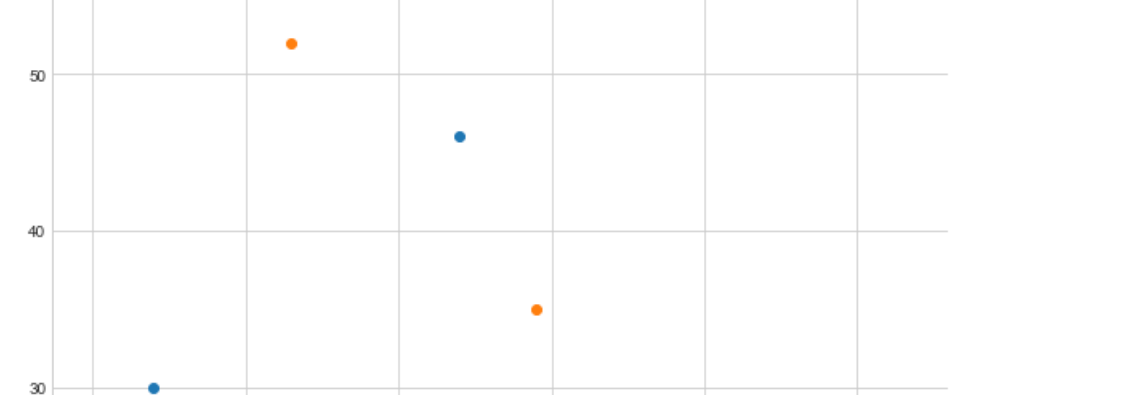
In [76]:

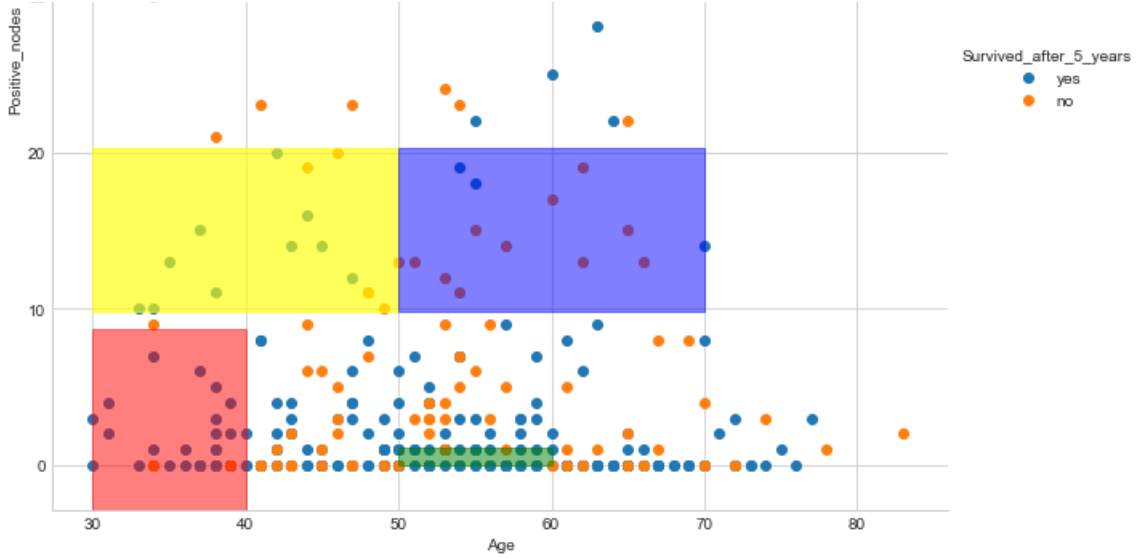
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_years"]
hbdata["Survived_after_5_years"] = hbdata["Survived_after_5_years"].map({1:"yes", 2:"no"})

sns.set_style("whitegrid");
sns.FacetGrid(hbdata, hue="Survived_after_5_years", size=8.5) \
   .map(plt.scatter, "Age", "Positive_nodes") \
   .add_legend();
plt.axvspan(xmin=30, xmax=40, ymin=0, ymax=.2, color='red',alpha = .5)
plt.axvspan(xmin=50, xmax=60, ymin=.05, ymax=.07, color='green', alpha= .5)
plt.axvspan(xmin=30, xmax=50, ymin=.22, ymax=.4, color='yellow', alpha= .65)
plt.axvspan(xmin=50, xmax=70, ymin=.22, ymax=.4, color='blue', alpha= .5)

plt.show();
```






# Observations:

1. Difficult to conclude things with surety however in the grid the leftmost and bottomost grid suggest compartively more Blue points compared to yellow, suggesting younger with lesser positive nodes had tendency for surviving much longer(more than 5 years).Check red color rectangle

   There are two more points that but my no means can say that this is some concrete conclusion.Thanks to grid feature i can see at two more places density of blue points is more.

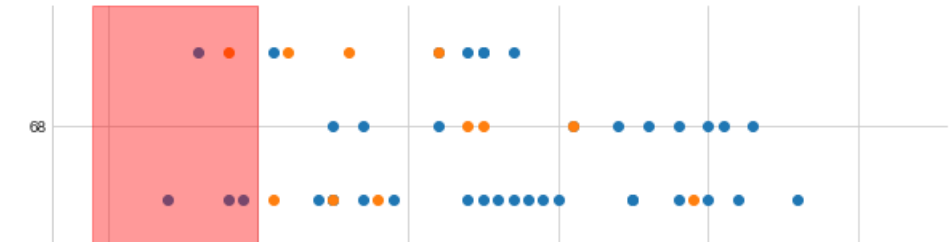
1. The bottomost third grid from left the line with zero positive nodes and age group between 50 to 60 filled with mostly blue line thus we can roughly conclude middle aged between with zero positive nodes lived much longer. Check the green color rectangle
2. Another very faint conclusion is from two grids falling between 10-20 positive nodes and age from 30-50 years there again the frequency of blue dots is more (check yellow rectangle) however the number of data points are fewer or could be there may be lots of data point completely overlapping so we are not able to see but with again can say people within 30-50 year age group with 10-20 positive nodes survived longer at the same time would like to point out i would like to study more through various other plots/graphs to further conclude this point.
3. Also as age is increasing (i.e 50-70 years) and positive nodes also increasing (10-20) the number of yellow dots is increasing and thus people are not living for more than 5 years in such cases. Check the blue rectangle

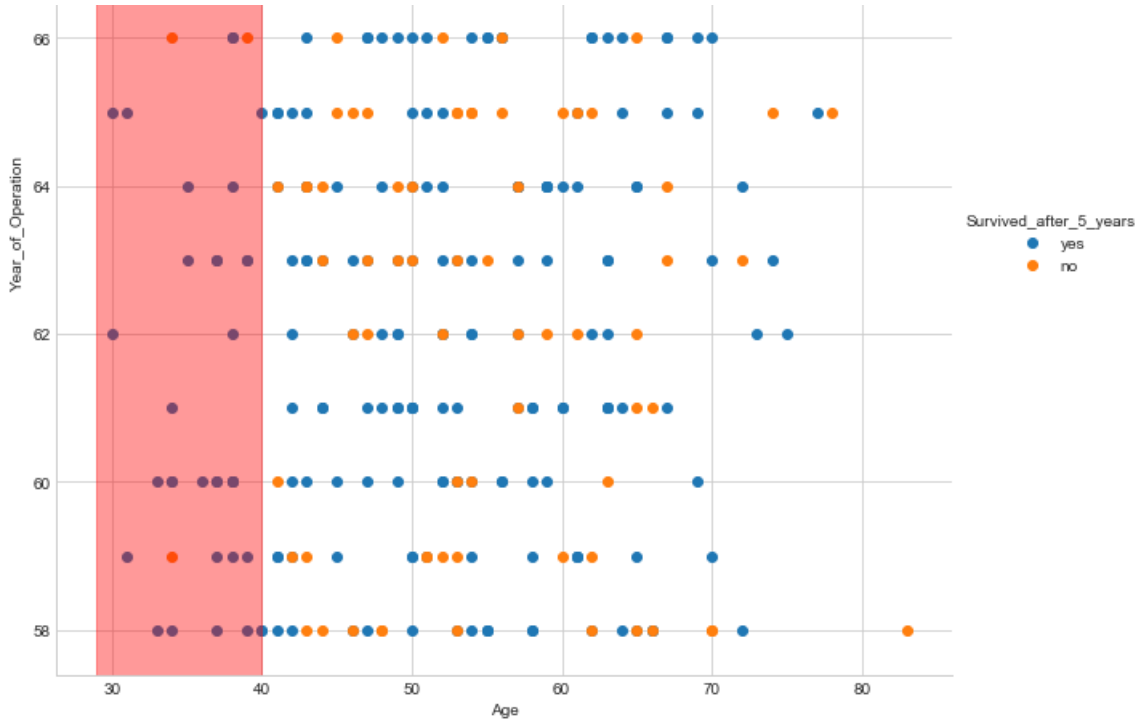
# Scatter Plot Year of Operation and Age (Plot 3)

In [48]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_years"]
hbdata["Survived_after_5_years"] = hbdata["Survived_after_5_years"].map({1:"yes", 2:"no"})
sns.set_style("whitegrid");
sns.FacetGrid(hbdata, hue="Survived_after_5_years", size=8.5) \
   .map(plt.scatter, "Age", "Year_of_Operation") \
   .add_legend();
plt.axvspan (xmin = 29, xmax = 40, color = 'red', alpha = .4 )
plt.show();
```






# Observations:

- Now if we see plot between Age and Year of Operation would like to point out we might be able to make some sense but the amount of lurking variables etc would be so many it would be foolishness to conclude anything but for fun sake lets continue.

***what we observe:***

1) For all the dots between 30-40 age group for all years number of blue dot is more compared to yellow(seems like 20-25 blue and 4-5 yellow). So we observe younger people tend to live longer however we dont know may be younger people having lesser positive nodes and thus living longer or is it even if they are having more positive nodes and still living longer? We have to see other graphs to come to any conclusion.Check red rectangle

2) There might be few more observation but i would say they are very futile and we should not "conclude" anything on the basis of "Age" and "year of operation"

**#**

We had already made all the possible scatter plot 3C2 = 3!/2! x 1! = 3

**# Overall Conclusion based on all the Scatter Plots:**

**1. It is very clear people with lesser positive nodes and lesser age tends to live more**

**2. People with lesser positive nodes irrespective of whatever age also tends to live but would be bit early to make concrete conslusion._**

**3. People with age more than 50+ and positive nodes between range 10-20 tends not to live longer than 5 years.**

End of scatter plot

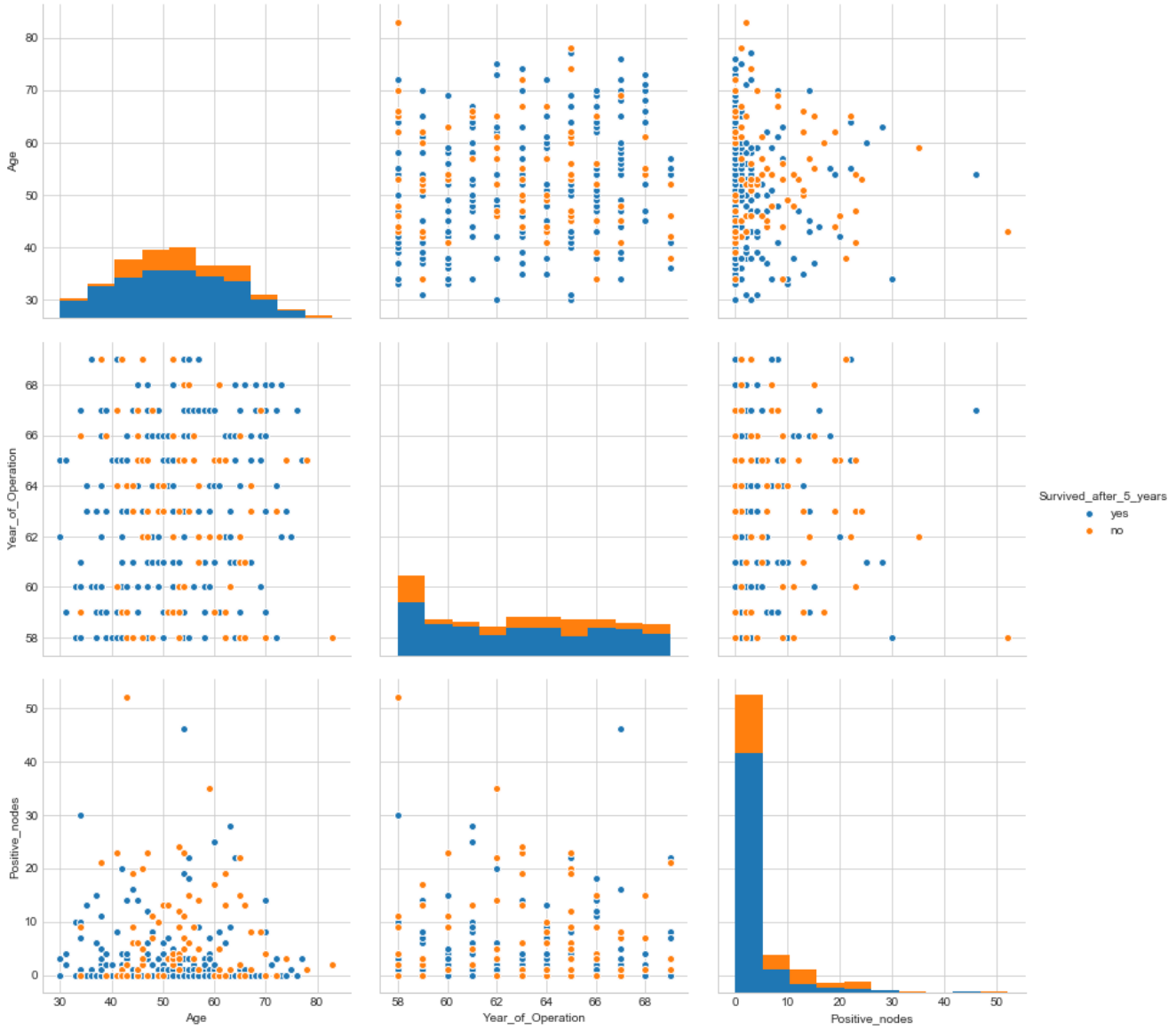
## Though we have seen all the possible scatter plot but we can also use "pair plot" command to draw all the plots in one shot.

***Please find the same below***

In [11]:

```
plt.close();
sns.set_style("whitegrid");
sns.pairplot(hbdata, hue="Survived_after_5_years", size=4);
```

```
plt.show()
```



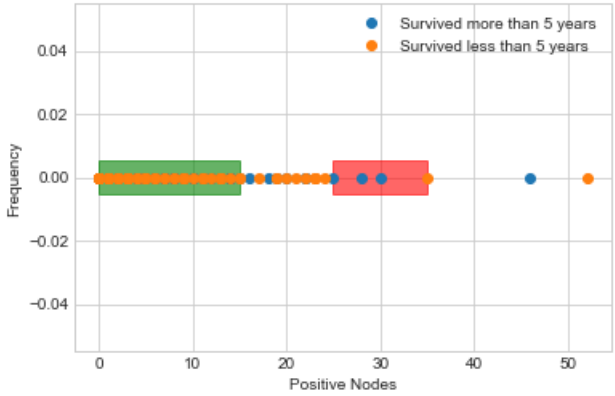

# Lets do some 1D analysis

In [71]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_years"]
hbdata["Survived_after_5_years"] = hbdata["Survived_after_5_years"].map({1:"yes", 2:"no"})
survived_longer = hbdata.loc[hbdata['Survived_after_5_years'] == "yes"]
survived_lesser = hbdata.loc[hbdata['Survived_after_5_years'] == "no"]

plt.plot(survived_longer["Positive_nodes"],np.zeros_like(survived_longer['Positive_nodes']), 'o',\
         label = "Survived more than 5 years")
plt.plot(survived_lesser["Positive_nodes"],np.zeros_like(survived_lesser['Positive_nodes']), 'o',
\
         label = "Survived less than 5 years")
plt.xlabel("Positive Nodes")
plt.ylabel("Frequency")
plt.axvspan (xmin = 0, xmax = 15, ymin = .45 , ymax = .55, color = "green", alpha = .6)
plt.axvspan (xmin = 25, xmax = 35, ymin = .45 , ymax = .55, color = "red", alpha = .6)
plt.legend()
plt.show()
```



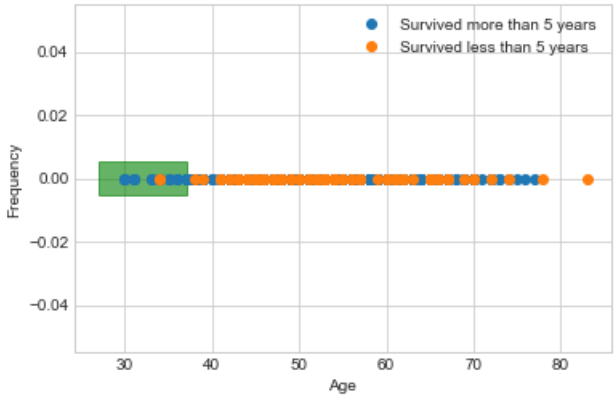

# Observation:

We see density of yellow points more between zero and 15 positive Nodes (Check green rectangle). After around 25 positive nodes we see more blue. (check red rectangle) dots.However we still dont know what are the counts of these dots but we can roughly conclude when postive nodes are less people tends to live longer compared when postive nodes is more than 25. For limitations of how many numbers of yellow n blue point lets look at histogram. Before that lets do one more 1 D analysis on the basis of "Age"

In [70]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_years"]
hbdata["Survived_after_5_years"] = hbdata["Survived_after_5_years"].map({1:"yes",2:"no"})
survived_longer = hbdata.loc[hbdata['Survived_after_5_years'] == "yes"]
survived_lesser = hbdata.loc[hbdata['Survived_after_5_years'] == "no"]
plt.plot(survived_longer["Age"],np.zeros_like(survived_longer['Age']), 'o', label = "Survived more
than 5 years")
plt.plot(survived_lesser["Age"],np.zeros_like(survived_lesser['Age']), 'o', label = "Survived less
than 5 years")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.legend()
plt.axvspan (xmin = 27,xmax = 37, ymin = .45, ymax = .55, alpha = .6, color = "green" )
plt.show()
```




# Observations:

1. All we can say is the density of blue points is more between 27-37 and somewhere around 70 years age group.Check the green rectangle

    - Howeverfrom this graph there is no info on postive nodes it could be the data that we have of younger people is of those with more positive nodes
    - Anyways all we can say from this graph people with age group around 25-30 and around 70-75 lived for longer years

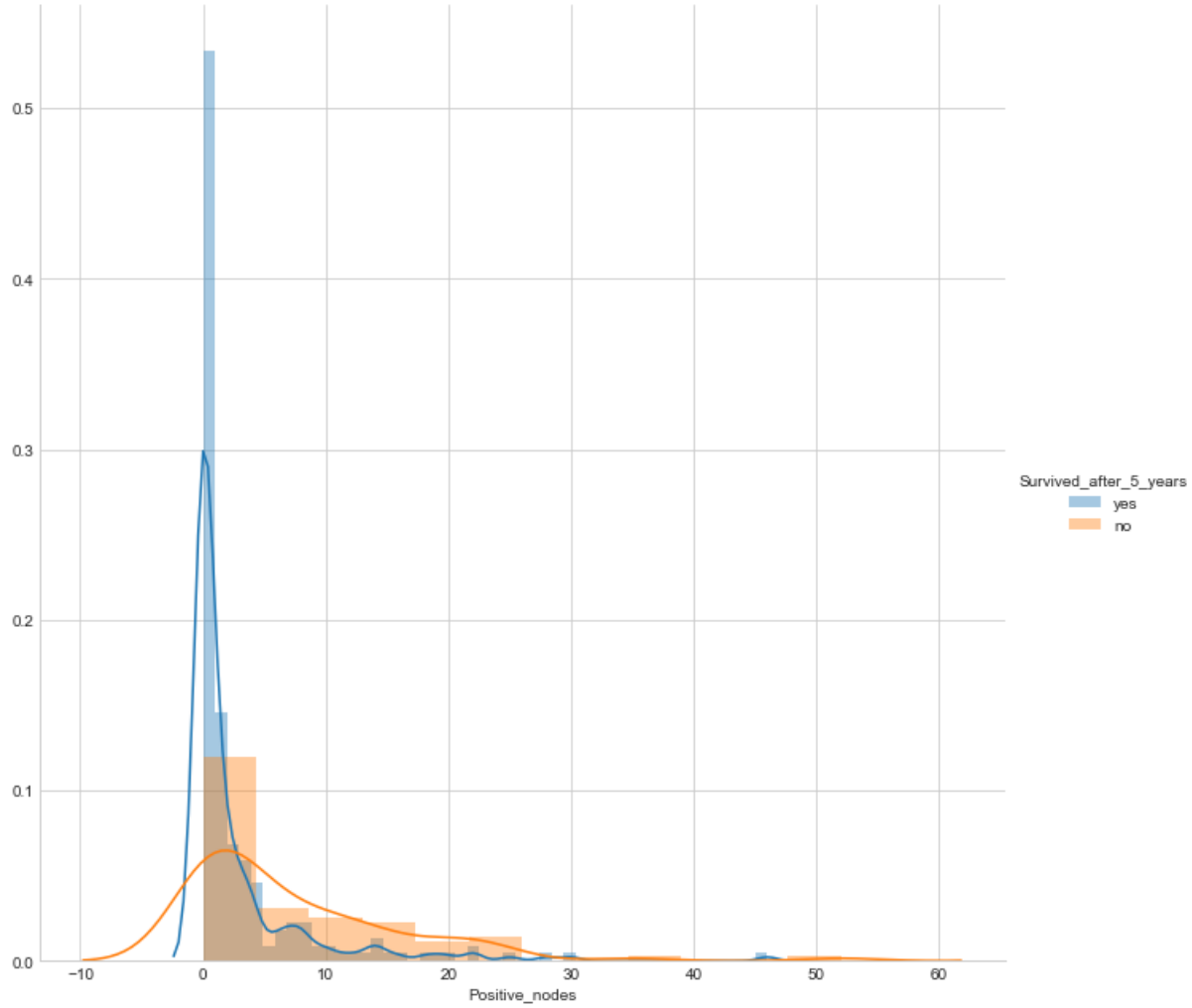
End of 1D Analysis

# Lets do histogram for above both graph

In [78]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.simplefilter("ignore")


hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_years"]
hbdata["Survived_after_5_years"] = hbdata["Survived_after_5_years"].map({1:"yes",2:"no"})
survived_longer = hbdata.loc[hbdata['Survived_after_5_years'] == "yes"]
survived_lesser = hbdata.loc[hbdata['Survived_after_5_years'] == "no"]
sns.FacetGrid(hbdata,hue = "Survived_after_5_years", size = 9).map(sns.distplot,"Positive_nodes").a
dd_legend();
plt.show();
```




# Observations

1. So again from this graph we can clearly say with lesser positive nodes people lives longer and as postive nodes increases tendency to live longer reduces
2. The maximum area of blue line is within limits of positive nodes that of little less than 0 and 5. So maximum people having positive nodes within positive nodes range limit from 0 to 5 .

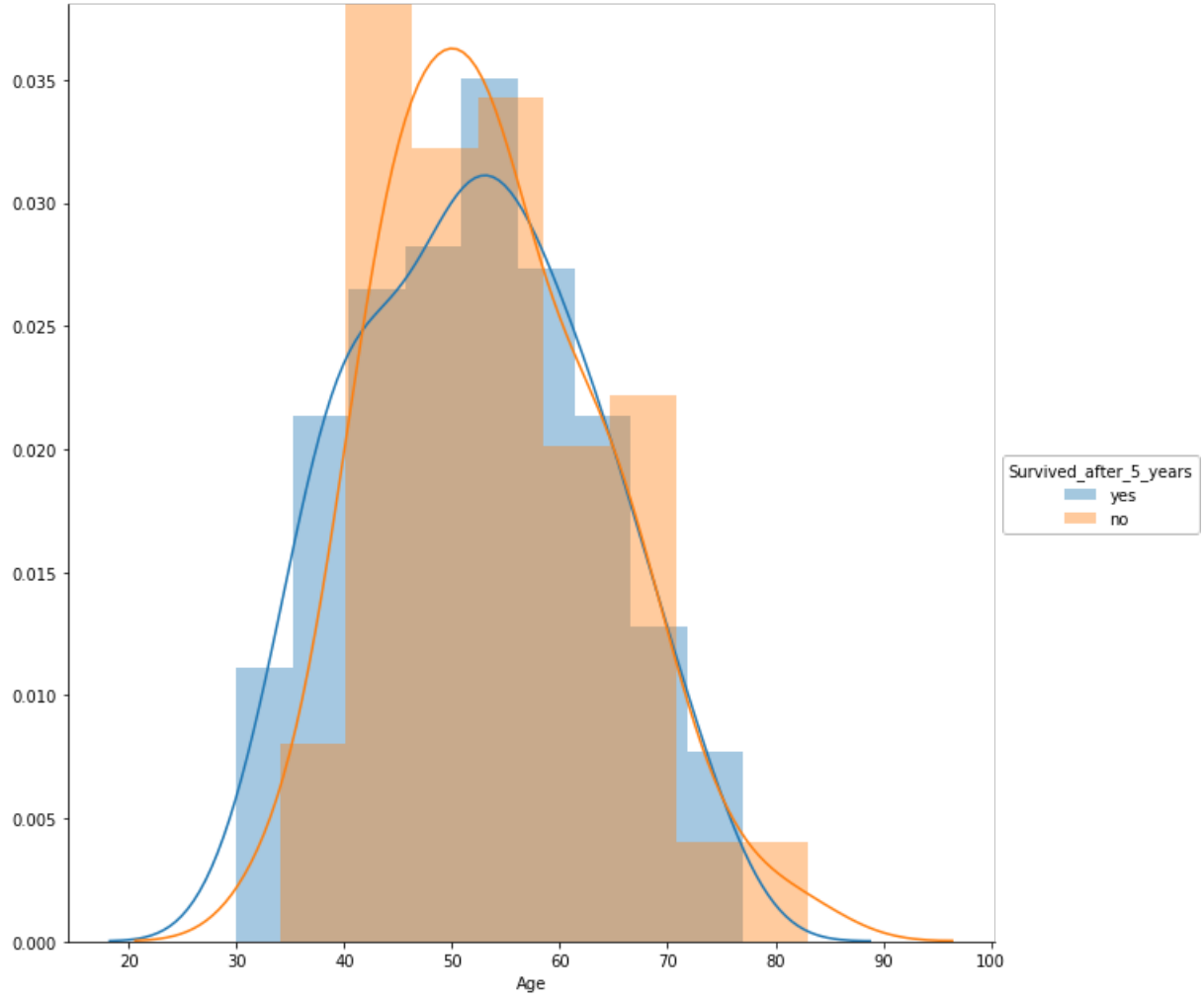
We might be able to tell exact % once we do CDF but for now we are happy with this conclusion

We might be able to tell exact % once we do CDF but for now we are happy with this conclusion

In [13]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.simplefilter("ignore")


hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_years"]
hbdata["Survived_after_5_years"] = hbdata["Survived_after_5_years"].map({1:"yes",2:"no"})
survived_longer = hbdata.loc[hbdata['Survived_after_5_years'] == "yes"]
survived_lesser = hbdata.loc[hbdata['Survived_after_5_years'] == "no"]
sns.FacetGrid(hbdata,hue = "Survived_after_5_years", size = 9).map(sns.distplot,"Age").add_legend()
;
plt.show();
```




# Observation:

Again as concluded earlier it wont be an good idea to come to any conclusion based alone on this graph as it is higly overlapped

End of Histogram

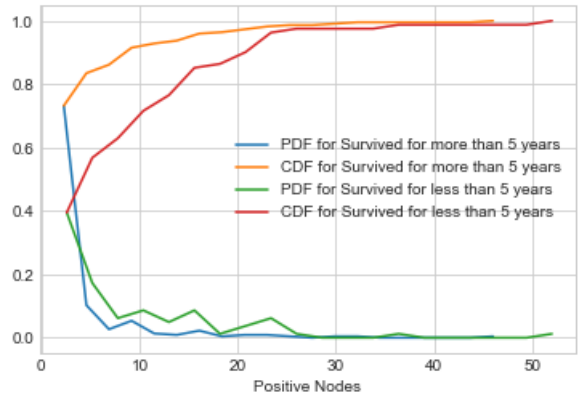
# Lets do PDF (probability Density Fuction) and CDF (Cummulative Density Function

In [95]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_Notes","Survived_after_5_Years"]
hbdata["Survived_after_5_Years"]=hbdata["Survived_after_5_Years"].map({1:"yes",2:"no"})
survived_longer = hbdata.loc[hbdata["Survived_after_5_Years"]=="yes"]
survived_lesser = hbdata.loc[hbdata["Survived_after_5_Years"]=="no"]
counts, bin_edges = np.histogram(survived_longer['Positive_Notes'], bins=20,density = True)
counts2, bin_edges2 = np.histogram(survived_lesser['Positive_Notes'], bins=20,density = True)

pdf = counts/(sum(counts))
pdf2 = counts2/(sum(counts2))
cdf = np.cumsum(pdf)
cdf2 = np.cumsum(pdf2)
plt.plot(bin_edges[1:],pdf,label = "PDF for Survived for more than 5 years");
plt.plot(bin_edges[1:], cdf, label = "CDF for Survived for more than 5 years")
plt.plot(bin_edges2[1:],pdf2,label = "PDF for Survived for less than 5 years");
plt.plot(bin_edges2[1:], cdf2, label = "CDF for Survived for less than 5 years")
plt.xlabel("Positive Nodes")
plt.legend()
plt.show();
```




# Observations 1

1. The below blue lines points to PDF i.e. probability density functions and orange line as CDF i.e. cummulative density function.Blue starts from around .7 mark and positive nodes being around 5. So we can say 70% people those survived more than 5 years had positive nodes less than 5
2. similarly when we see the organge line corresponding to positive nodes point of 10 it almost points somewhere between .85 or so. So almost 85% of people who survived more than 5 years has positive nodes less than 10

# Observation 2

However also from other red and green line we can make following conclusions:

1. In plot we see from positive nodes 5 till 27 or so green line is constantly above blue line. So probability of surviving less than 5 years (depicted by green line) is more than surviving more than 5 years (depicted by blue line).

# Lets do on the basis of Age

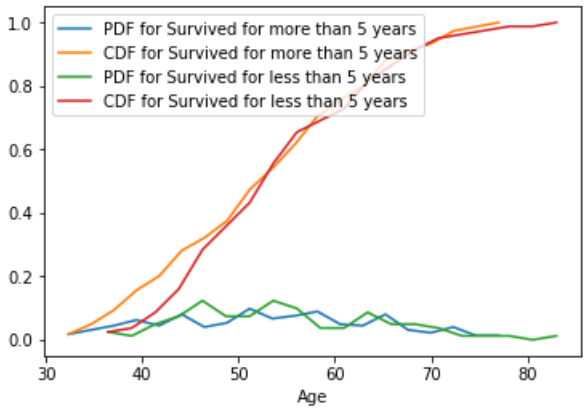
In [2]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_Nodes","Survived_after_5_Years"]
hbdata["Survived_after_5_Years"]=hbdata["Survived_after_5_Years"].map({1:"yes",2:"no"})
```

```
survived_longer = hbdata.loc[hbdata["Survived_after_5_Years"]=="yes"]
survived_lesser = hbdata.loc[hbdata["Survived_after_5_Years"]=="no"]
counts, bin_edges = np.histogram(survived_longer['Age'], bins=20,
                                 density = True)
counts2, bin_edges2 = np.histogram(survived_lesser['Age'], bins=20,density = True)

pdf = counts/(sum(counts))
pdf2 = counts2/(sum(counts2))
cdf = np.cumsum(pdf)
cdf2 = np.cumsum(pdf2)
plt.plot(bin_edges[1:],pdf,label = "PDF for Survived for more than 5 years");
plt.plot(bin_edges[1:], cdf, label = "CDF for Survived for more than 5 years")
plt.plot(bin_edges2[1:],pdf2,label = "PDF for Survived for less than 5 years");
plt.plot(bin_edges2[1:], cdf2, label = "CDF for Survived for less than 5 years")
plt.xlabel("Age")
plt.legend()
plt.show()
```




# Conclusions

OK now we get some clearer picture regarding "age" variablle.

1. The Blue line is pretty much constant through out means probability of people living more or less than 5 years in pretty much same irrespective of age. Thus we should NOT BE making any conclusions on the basis of age.
2. The same trend can be seen with respect to green line it is pretty much constant and straight and thus we should not be concluding on the basis of age.
3. There is natural bend downwards for both green and blue line as age increases with is pretty natural as people become old there tendeny to live more than 5 years is reducing and thus a donward bend trend can be observed

**So moving forward i wont be doing much analysis for age much. We are left with Box and Violin plot. Lets dig in that**

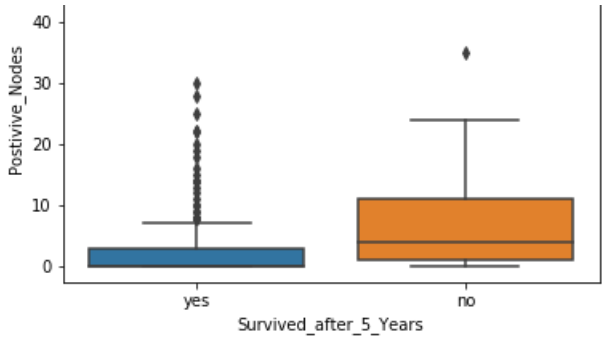
# Lets do Box Plot

## Box Plot on the basis of Positive Nodes

In [24]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_Years"]
hbdata["Survived_after_5_Years"]=hbdata["Survived_after_5_Years"].map({1:"yes",2:"no"})
sns.boxplot(x='Survived_after_5_Years',y='Positive_nodes', data=hbdata)
plt.show()
```




# Conclusions

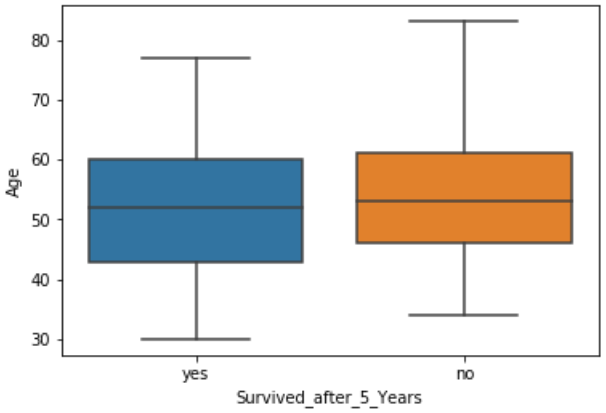
1. So we can see the blue box which gives us clear picture of quantile tells us almost 75% of people who lived more than 5 years has positive nodes less than 4 or so .
2. Almost 75% of people who did not lived longer than five years had positive nodes less than 12 or so

## Box Plot on the basis of Age

In [27]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_Years"]
hbdata["Survived_after_5_Years"]=hbdata["Survived_after_5_Years"].map({1:"yes",2:"no"})
sns.boxplot(x='Survived_after_5_Years',y='Age', data=hbdata)
plt.show()
```




# Conclusion

Age is not such a good "variable" to conclude much

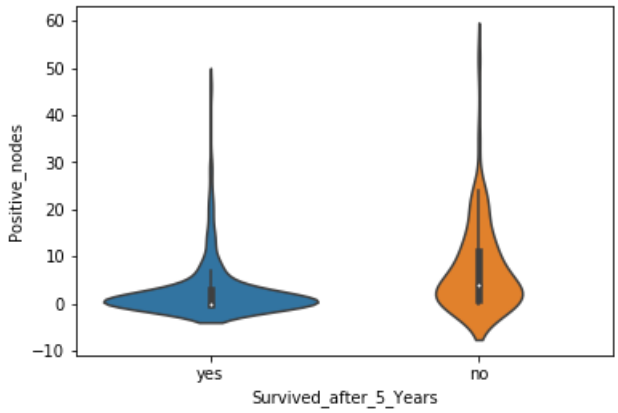
# Lets do Violin Plot

In [28]:

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

hbdata = pd.read_csv(r"D:\AppliedAI\Homework n Assignments\habermans-survival-data-
set\haberman.csv")
hbdata.columns = ["Age","Year_of_Operation","Positive_nodes","Survived_after_5_Years"]
hbdata["Survived_after_5_Years"]=hbdata["Survived_after_5_Years"].map({1:"yes",2:"no"})
```

```
sns.violinplot(x="Survived_after_5_Years", y="Positive_nodes", data=hbdata, size=8)
plt.show()
```




# Conclusion

1. From this we can conclude that it might be true that for both people staying more than five years and people less than five years had lesser number of Positive Nodes but the blue violin (i.e.) that of people staying more than five years is more fatter when positive nodes is near zero.
2. Thus we can conclude people with positive nodes near zero has "more" probability of staying more than 5 years compared to having positive nodes more than zero. Further the violin for people living less than 5 years keep still has some width when positive nodes is large but the blue violin width almost vanishes when postive nodes increases thus further solidyfying our conclusion