INSAID

INTERNATIONAL SCHOOL OF AI AND DATA SCIENCE

Table of Contents

1 Problem Statement

To analyse and find out the possiblity of patients who will survive beyond 5 years after surgery based on the attributes like Age, Operation Year and Axillary nodes.

2 Data Loading and Description

Data Set Information: 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. Attribute Information:

1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute) -- 1 = the patient survived 5 years or longer -- 2 = the patient died within 5 year

2.1 Importing libraries

In [ ]:

```
import numpy as np
import pandas_profiling
import pandas as pd
pd.set_option('mode.chained_assignment', None)      # To suppress pandas warning
pd.set_option('display.max_colwidth', -1)           # To display all the data in
pd.options.display.max_columns = 50                 # To display every column of

import warnings
warnings.filterwarnings('ignore')                   # To suppress all the warnin

import matplotlib.pyplot as plt
%matplotlib inline

import seaborn as sns
sns.set(style='whitegrid', font_scale=1.3, color_codes=True)     # To apply seab

# Making plotly specific imports
# These imports are necessary to use plotly offline without signing in to their
```

In [2]:

```
# Install and update plotly using this command to the latest version (higher than
#!pip install plotly --upgrade
```

In [3]:

```
# Install chart_studio, please use this command.
#!pip install chart-studio
```

2.2 Loading dataset

In [3]:

```
excel_df = pd.read_csv('https://raw.githubusercontent.com/insaid2018/Term-1/maste
```

3 Data Profiling

3.1 Get a basic information on Dataset

In [4]:

```
excel_df.shape                                             # It will return the coun
```

Out[4]:

```
(306, 4)
```

Above result represents that we have 306 rows and 4 columns in our dataset

In [5]:

```
excel_df.head()                                    # It will return first 5
```

Out[5]:

	Age	Years_of_operation	Pos_axillary_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 [6]:

```
excel_df.info();                                  # This will give Index, Dat
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306 entries, 0 to 305
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 306 non-null    int64
 1   Years_of_operation  306 non-null    int64
 2   Pos_axillary_nodes  306 non-null    int64
 3   Status              306 non-null    int64
dtypes: int64(4)
memory usage: 9.7 KB
```

We have observed the below points

- No missing values
- All the columns are of the integer data type.
- **Status -** To get a better understanding of Survival status, it has to be converted into **Yes** or **No** category
  - The value of '1' will be replaced with 'Yes' to denote the patient has survived more than 5 years.
  - The value of '2' will be replaced with 'No' to denote the patient has survived less than 5 years.
- **Years of operation -** As we can see all the surgeries are conducted in between 1958 and 1970, we can add the value '1900' to this column to make it more clear

3.2 Pre Profiling

In [8]:

```
# To install pandas profiling please run this command.

#!pip install --user pandas-profiling --upgrade
#!pip install -q datascience
#!pip install --user --upgrade pandas-profiling
```

In [9]:

```
profile = excel_df.profile_report(title='Pandas Profiling before Data Preprocess
```

- Convert the 'status' column values into categorical datatype
- Sum each value of 'Years_of_operation' column with a integer value of '1900'

In [7]:

```
excel_df['Status'] = excel_df['Status'].map({1:'Yes', 2:'No'})
excel_df['Years_of_operation'] = excel_df.Years_of_operation + 1900;
excel_df.head()
```

Out[7]:

	Age	Years_of_operation	Pos_axillary_nodes	Status
0	30	1964	1	Yes
1	30	1962	3	Yes
2	30	1965	0	Yes
3	31	1959	2	Yes
4	31	1965	4	Yes

3.3 Post Profiling

In [8]:

```
excel_df.describe()
```

Out[8]:

	Age	Years_of_operation	Pos_axillary_nodes
count	306.000000	306.000000	306.000000
mean	52.457516	1962.852941	4.026144
std	10.803452	3.249405	7.189654
min	30.000000	1958.000000	0.000000
25%	44.000000	1960.000000	0.000000
50%	52.000000	1963.000000	1.000000
75%	60.750000	1965.750000	4.000000
max	83.000000	1969.000000	52.000000

In [14]:

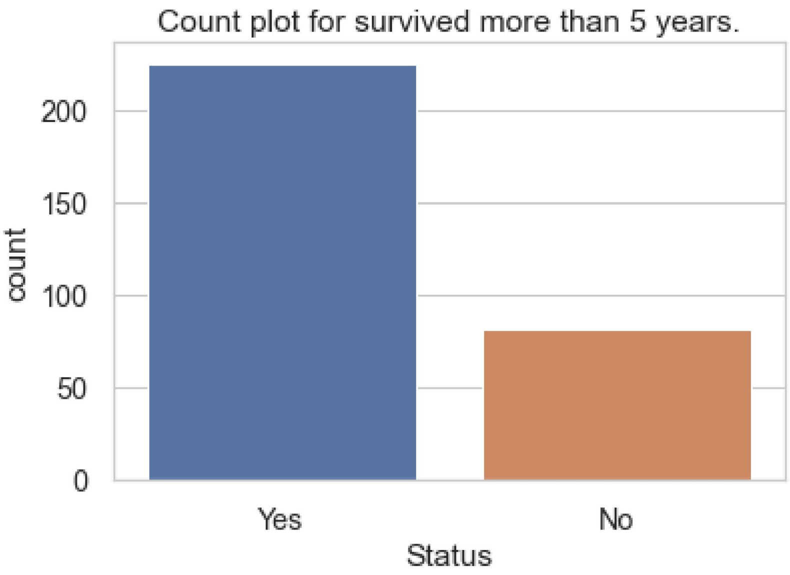
```
# To output pandas profiling report to an external html file.
# Saving the output as profiling_before_preprocessing.html

profile = excel_df.profile_report(title='Pandas Profiling before Data Preprocess
profile.to_file(output_file="profiling_before_preprocessing.html")
```

4 Questions

4.1 What is the overall survival rates?

```
In [8]: sns.countplot(x='Status', data=excel_df).set_title('Count plot for survived more
```

```
Out[8]: Text(0.5, 1.0, 'Count plot for survived more than 5 years.')
```



If we want to know the exact count who are all survived more than 5 years.

Using Groupby of Status column

```
In [19]: excel_df.groupby(['Status'])['Status'].count()
```

```
Out[19]: Status
         No      81
         Yes    225
         Name: Status, dtype: int64
```

As per the above count, we can say 225 patients has survived more than 5 years after the surgery and 81 patients survival rate is below 5 years.

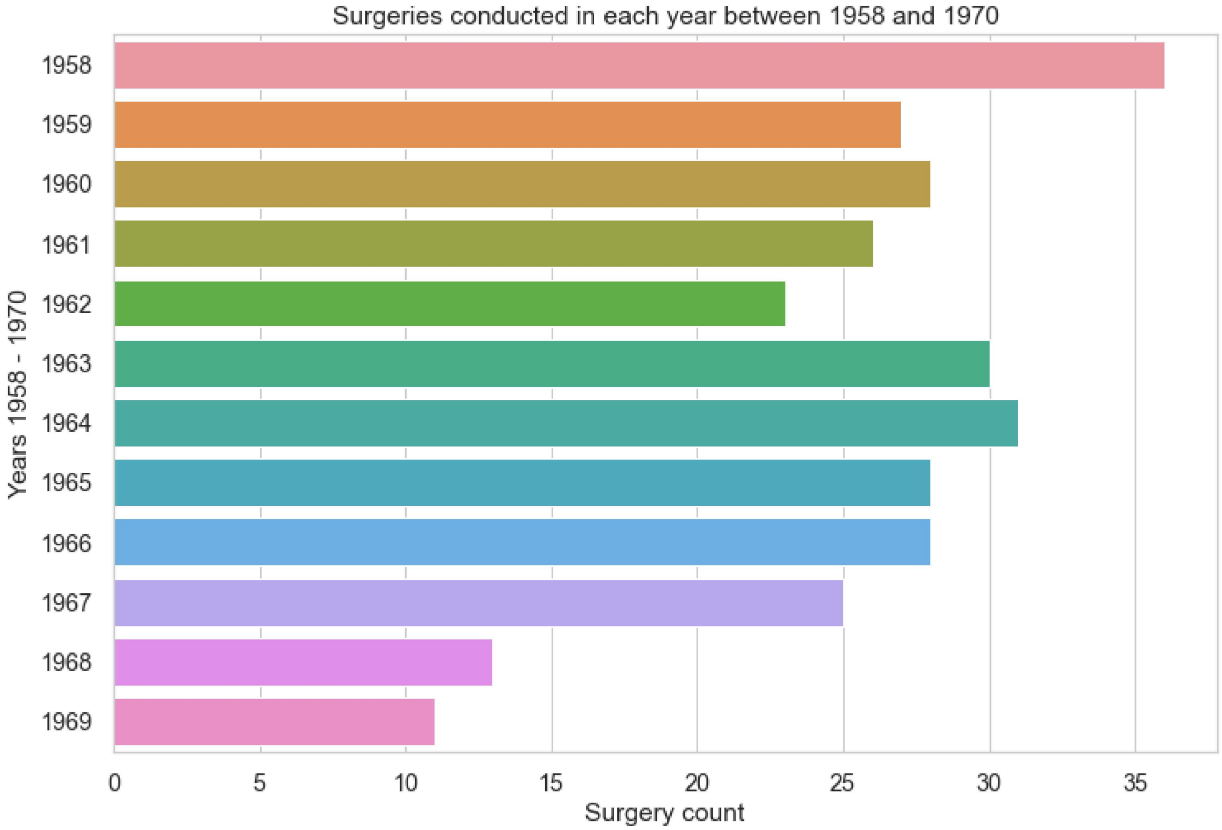
4.2 Which year has maximum number of surgeries?

In [15]:

```
plt.figure(figsize=(12,8))
sns.countplot(y='Years_of_operation', data=excel_df).set_title('Surgeries conduc
plt.ylabel('Years 1958 - 1970')
plt.xlabel('Surgery count')
```

Out[15]:

```
Text(0.5, 0, 'Surgery count')
```



To know the exact count, we use the groupby for Years_of_operation.

```
In [23]: #excel_df.groupby(['Years_of_operation'])['Years_of_operation'].count()
         excel_df.groupby(['Years_of_operation'])['Years_of_operation'].count().sort_valu
```

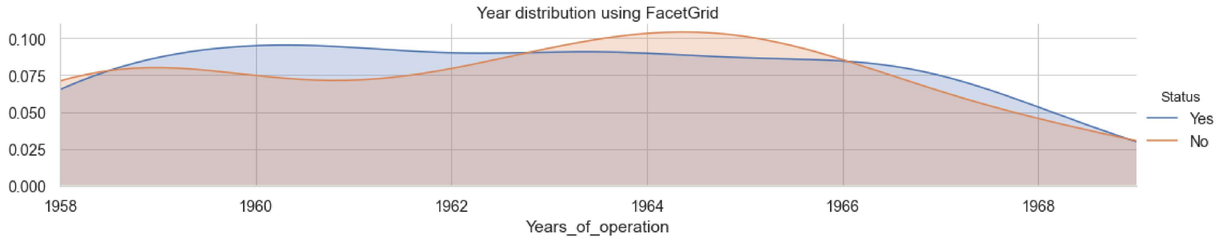
```
Out[23]: Years_of_operation
         1958    36
         1964    31
         1963    30
         1966    28
         1965    28
         1960    28
         1959    27
         1961    26
         1967    25
         1962    23
         1968    13
         1969    11
         Name: Years_of_operation, dtype: int64
```

As shown in the above graph, Maximum surgeries recorded in the year of 1958, followed by 1964.

4.3 Which year has lowest success rate in surgeries ?

```
In [23]: as_fig = sns.FacetGrid(excel_df,hue='Status',aspect=5)

         as_fig.map(sns.kdeplot,'Years_of_operation',shade=True)

         firstyear = excel_df['Years_of_operation'].min()

         Lastyear = excel_df['Years_of_operation'].max()

         as_fig.set(xlim=(firstyear,Lastyear))

         as_fig.add_legend()
         plt.title('Year distribution using FacetGrid')
```

```
Out[23]: Text(0.5, 1.0, 'Year distribution using FacetGrid')
```



Major area of graph has been overlapped which tells that survival chances of patients cannot be decided using only this **Years_of_operation** parameter.

Between the year 1963 and 1966, We can see a rise in the graph which tells that there were more unsuccessful operations.

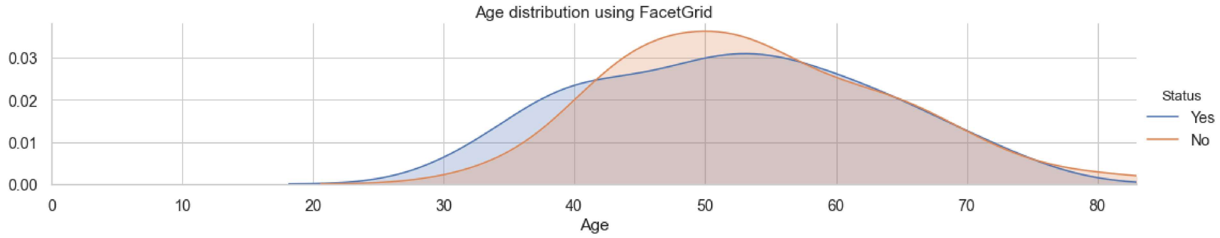
4.4 Does the age factor decide the survival chance?

In [17]:

```
as_fig = sns.FacetGrid(excel_df,hue='Status',aspect=5)

as_fig.map(sns.kdeplot,'Age',shade=True)

oldest = excel_df['Age'].max()

as_fig.set(xlim=(0,oldest))

as_fig.add_legend()
plt.title('Age distribution using FacetGrid')
```

Out[17]:

```
Text(0.5, 1.0, 'Age distribution using FacetGrid')
```

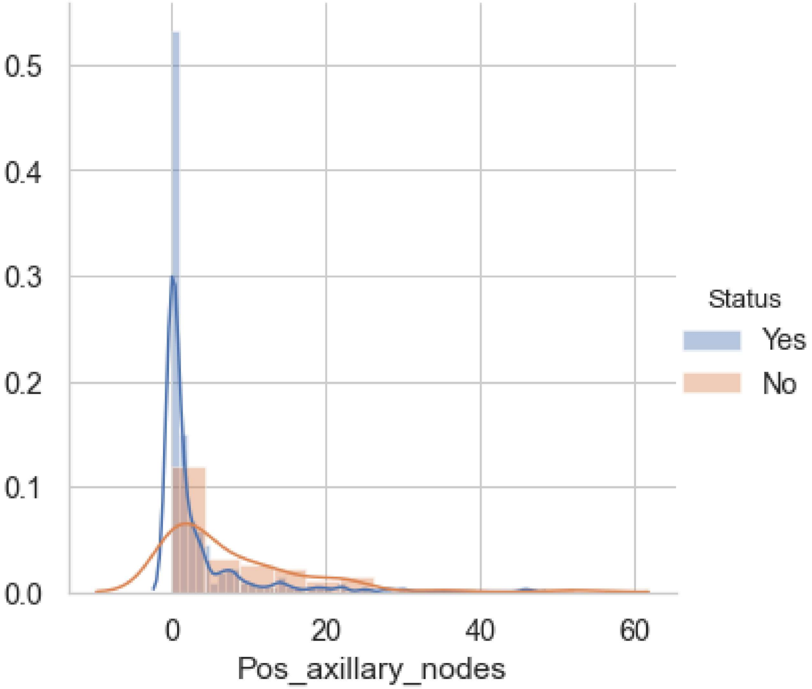


From the above graph, Overlapping area shows that the survival chances for patients are almost similiar for all the age category. Hence we cannot decide the survival chances of a patient using only the age as a parameter. Lets see the below observations

1. Patients age between 25 and 40 are slightly having higher chance to survive after 5 years.
2. Patients age which are more than 40 and less than 55 are having less chances to survive after 5 years.
3. Patients whose age are between 60 and 75 are having equal chances of surviving and not surviving, while the survival chances are less after the age of 80.

4.5 Does the axillary nodes decide the survival chance?

In [18]:
```
sns.FacetGrid(excel_df,hue='Status',height = 5)\
 .map(sns.distplot,"Pos_axillary_nodes")\
 . add_legend();
plt.show()
```



Patients with less nodes are more likely to survive.
From the above graph, we can clearly see the patients who has less than 2 nodes are having more chances to survive more than 5 years.

In [11]:
```
LongSurvival = excel_df[excel_df['Status'] == 'Yes']

print("Long Surival\n")
print("Total count is " + str(LongSurvival['Pos_axillary_nodes'].count()) )
print("Mean - " + str(np.mean(LongSurvival['Pos_axillary_nodes'])))
print("Median - "+ str(np.median(LongSurvival['Pos_axillary_nodes'])))
print("Maximum node in Long Survival - "+ str(np.max(LongSurvival['Pos_axillary_
print("\n**********************************************\n")

ShortSurvival = excel_df[excel_df['Status'] == 'No']

print("Short Surival\n")
print("Total count is " + str(ShortSurvival['Pos_axillary_nodes'].count()) )
print("Mean - " + str(np.mean(ShortSurvival['Pos_axillary_nodes'])))
print("Median - "+ str(np.median(ShortSurvival['Pos_axillary_nodes'])))
print("Minimum node in Short Survival - "+ str(np.min(ShortSurvival['Pos_axillary
```

```
Long Surival

Total count is 225
Mean - 2.7911111111111113
Median - 0.0
Maximum node in Long Survival - 46

**********************************************

Short Surival

Total count is 81
Mean - 7.45679012345679
Median - 4.0
Minimum node in Short Survival - 0
```

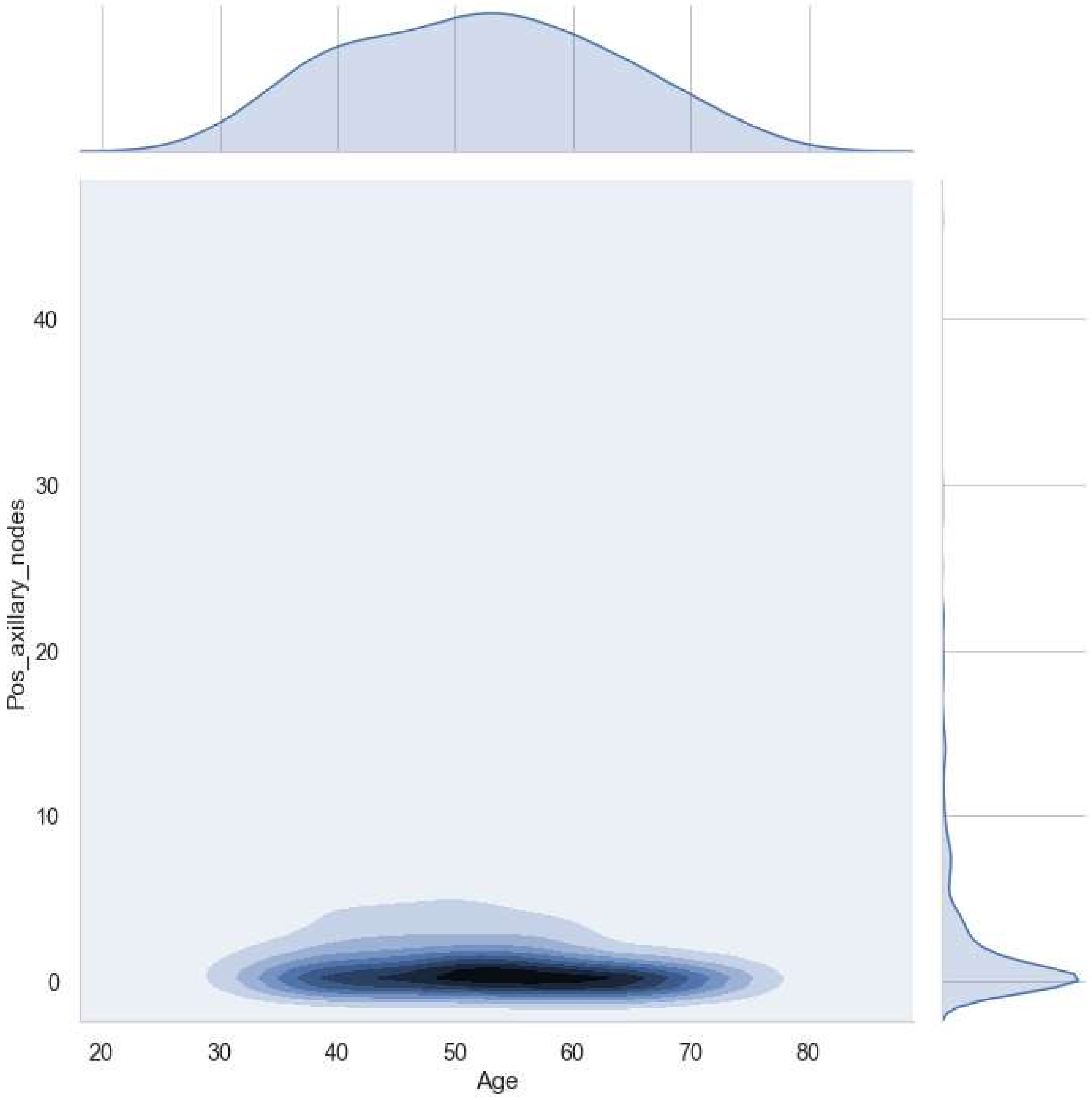
Based on the above code snippet, we have observed the below points

- **Long Survived -** Median value of "0" represents that majority of long survived patients are having "0" axillary nodes.
- **Short Survived -** Median value of "4.0" represents that majority of short survived patients are having "4" axillary nodes.

Note: Data also represents that few patients in the Long survival category has more than 40 nodes, also the patients in the short survival category has no positive nodes. This represents that there is some possibility of incorrect information present in the data.

4.6 Which age category has maximum numbers of Long Survival?

In [129]:
```
sns.jointplot(x = "Age", y = "Pos_axillary_nodes", data = LongSurvival, kind = "
plt.show()
```

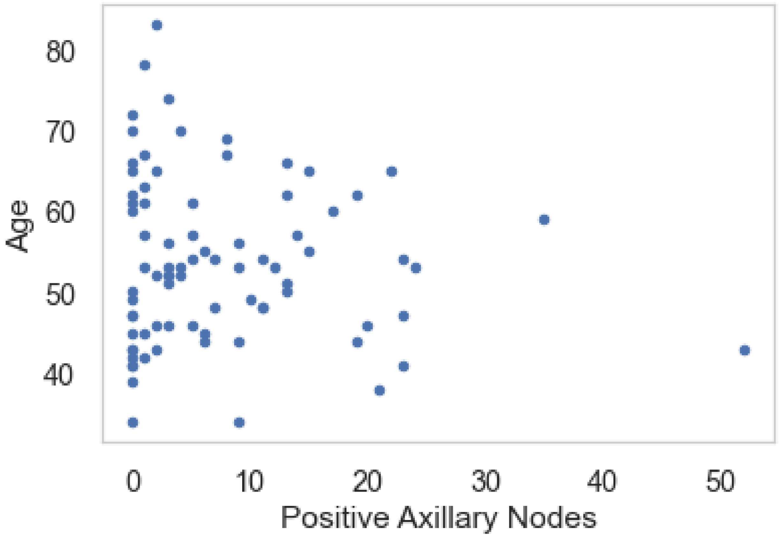


Above is the contour plot for long survival using age and axillary nodes, it is observed that density for long survival is more from the age range 47–60 and axillary nodes from 0–3.

4.7 What percentage of patients who had no positive axillary nodes died within 5 years?

In [37]:
```
ShortSurvival.plot(kind="scatter", x="Pos_axillary_nodes", y="Age")
plt.ylabel('Age')
plt.xlabel('Positive Axillary Nodes')
plt.grid()
plt.show()
```

```
'c' argument looks like a single numeric RGB or RGBA sequence, which should be
avoided as value-mapping will have precedence in case its length matches with
'x' & 'y'.  Please use a 2-D array with a single row if you really want to spec
ify the same RGB or RGBA value for all points.
```



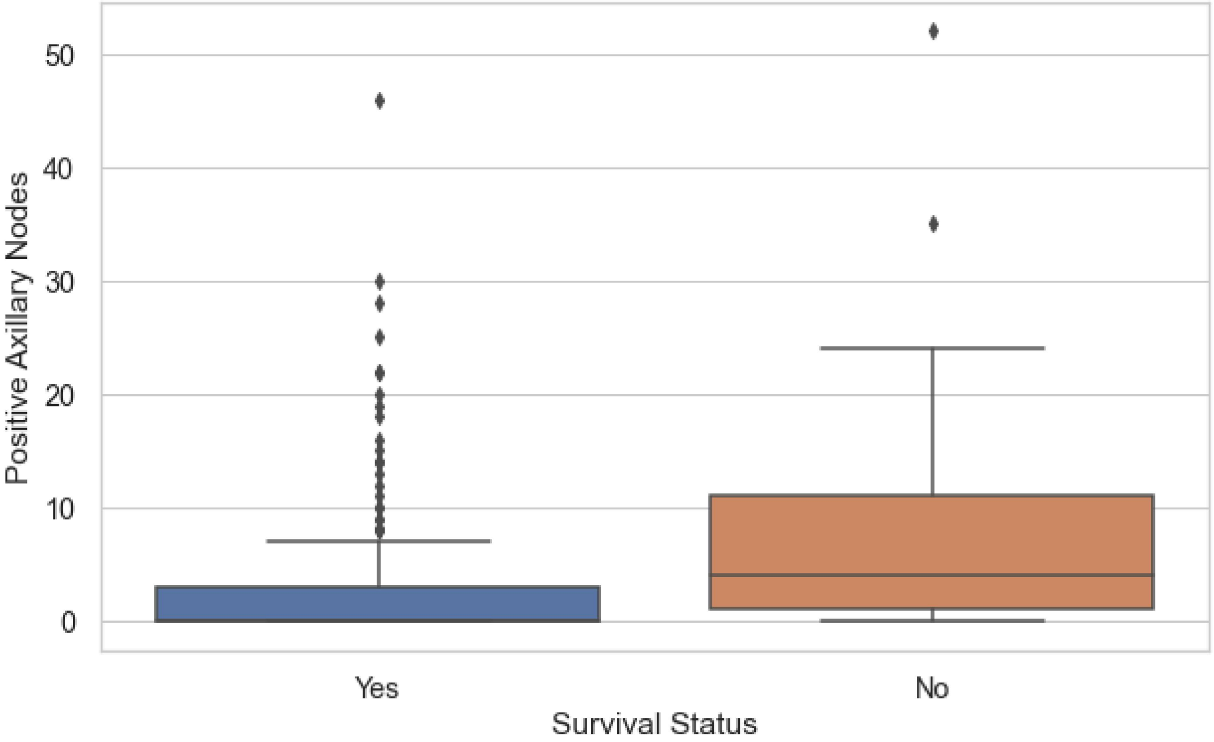
In [39]:
```
print(ShortSurvival[ShortSurvival['Pos_axillary_nodes'] <= 0].count())
zeronodespatients = ShortSurvival[ShortSurvival['Pos_axillary_nodes'] <= 0].Pos_
print("\nPatients with 0 nodes who were survived less than 5 years are '{}'%\n".
```

```
Age                   19
Years_of_operation    19
Pos_axillary_nodes    19
Status                19
dtype: int64

Patients with 0 nodes who were survived less than 5 years are '6.21'%
```

Lets see the Survival status using Box plot

In [82]:

```
plt.figure(figsize=(10,6))
sns.boxplot(x="Status",y="Pos_axillary_nodes", data=excel_df)
plt.ylabel('Positive Axillary Nodes')
plt.xlabel('Survival Status')
plt.show()
```



In the above box plot, following points needs to be considered

- 25th percentile and 50th percentile are nearly same for Long survive and threshold for it is 0 to 7.
- For short survival there are 50th percentile of nodes are nearly same as long survive 75th percentile. Threshold for the Short survival is 0 to 25 nodes
- Large percentage of patients who survived had 0 nodes. But there is a small percentage of 6.21 % patients who had no positive axillary nodes died within 5 years of operation, thus an absence of positive axillary nodes cannot always guarantee survival.
- Patients with more than 1 nodes are not likely to survive. More the number of nodes, lesser the survival chances.

5 Conclusion

- Patients with less than 35 years of age will survive more than 5 years

- Patients who has less nodes having high rate of survival. However, the data has shown that an absence of positive axillary nodes does not always guarantee survival. (Refer 4.7)
- Patients who had undergone surgery in the year between 1963 and 1966 having high number of less survival rate