Importing the Haberman dataset

In [51]:

```
#Om Ganeshay Namah
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
df = pd.read_csv(r'C:\Users\sandeep.chotalia\Documents\AppliedAI\Assignments\haberman.csv')
```

In [52]:

```
print(df)
```

```
     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
..   ...   ...    ...     ...
301   75    62      1       1
302   76    67      0       1
303   77    65      3       1
304   78    65      1       2
305   83    58      2       2

[306 rows x 4 columns]
```

Changing Column names to make it more meaningful

In [53]:

```
df.columns=['Patient Age','TreatmentYear','PositiveLymphNodes','Survivalstatus5years']
print(df)
```

```
     Patient Age  TreatmentYear  PositiveLymphNodes  Survivalstatus5years
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
..           ...            ...                 ...                   ...
301           75             62                   1                     1
302           76             67                   0                     1
303           77             65                   3                     1
304           78             65                   1                     2
305           83             58                   2                     2

[306 rows x 4 columns]
```

Number of DataPoints and Features

In [54]:

```
print(df.shape)
```

```
(306, 4)
```

There are 306 rows and 4 columns in the dataset haberman

Column Names in the dataset

In [55]:

```
print(df.columns)
```

```
Index(['Patient Age', 'TreatmentYear', 'PositiveLymphNodes',
       'Survivalstatus5years'],
      dtype='object')
```

From the above data we can see that there are 4 columns of type object which are 'age','year','nodes','status'

Explanation of the columns in the dataset 'age' = Age of Patient at time of operation of datatype numerical 'year' = Year of operation of datatype numerical 'nodes' = Number of positive axillary nodes detected of datatype numerical 'status' = Survival status of the patient 1 = the patient survived 5 years or longer Survival status of the patient 2 = the patient died within 5 years No missing attributes

In [61]:

```
print(df.describe())
```

```
       Patient Age  TreatmentYear  PositiveLymphNodes  Survivalstatus5years
count   306.000000     306.000000          306.000000            306.000000
mean     52.457516      62.852941            4.026144              1.264706
std      10.803452       3.249405            7.189654              0.441899
min      30.000000      58.000000            0.000000              1.000000
25%      44.000000      60.000000            0.000000              1.000000
50%      52.000000      63.000000            1.000000              1.000000
75%      60.750000      65.750000            4.000000              2.000000
max      83.000000      69.000000           52.000000              2.000000
```

The 'SurvivalStatus5years' has to be changed to categorical datatype as this is the main objective of our problem and we want to carry out our EDA based upon that.

In [64]:

```
status = {1:'yes',2:'no'}
df.Survivalstatus5years = [status[newval] for newval in df.Survivalstatus5years]
```

In [65]:

```
df["Survivalstatus5years"].value_counts()
```

Out[65]:

```
yes    225
no      81
Name: Survivalstatus5years, dtype: int64
```

In [66]:

```
print(df.iloc[:,-1].value_counts(normalize=True))
```

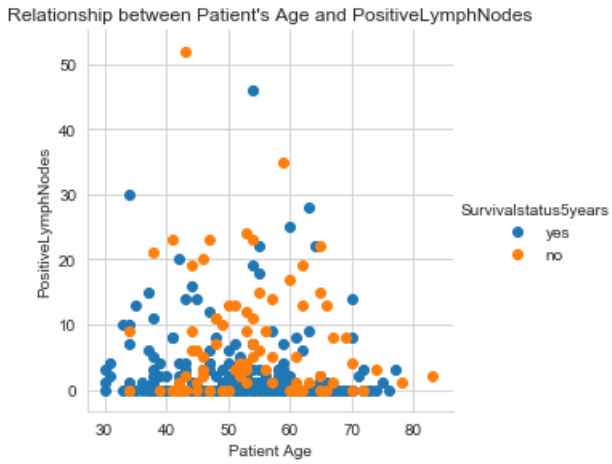
```
yes    0.735294
no     0.264706
Name: Survivalstatus5years, dtype: float64
```

Observations: 1) The minimum to maximum range of the Patients ranges from 30 years to 83 years. 2) The average age of the patients is about 52 years. 3) The maximum number of positive lymph nodes is 52, out of which 75% of the age group have less than 4 positive lymph nodes while 50% and 25% have less than 5 or no positive lymph nodes respectively. 4) It is an imbalanced dataset as 73% of the values are in the Survival side while only 26% are not in the survival side. So a big gap is observed between survival and non survival side, hence imbalanced dataset

In [101]:

```
#2-D Scatter plot with color-coding for survival status class between Patient's Age and PositiveLymphNo
des
```

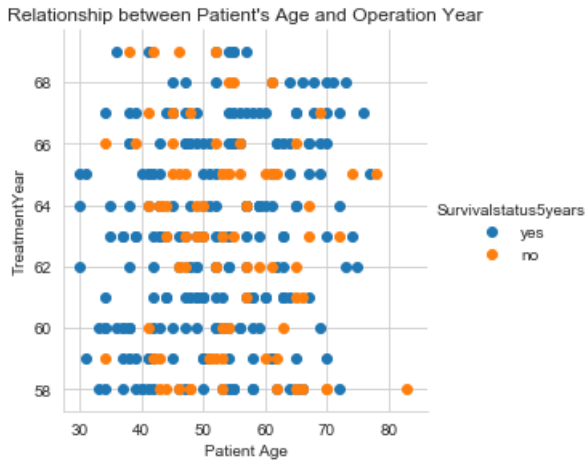
```
sns.set_style("whitegrid")
sns.FacetGrid(df, hue='Survivalstatus5years',height=4) \
   .map(plt.scatter,"Patient Age","PositiveLymphNodes") \
   .add_legend();
plt.title("Relationship between Patient's Age and PositiveLymphNodes")
plt.show();
```



Observation:From the above plot we can see that there are too many overlapping values between a Patient's age and Postive Nodes. So its very difficult to choose the relevant variable

In [104]:

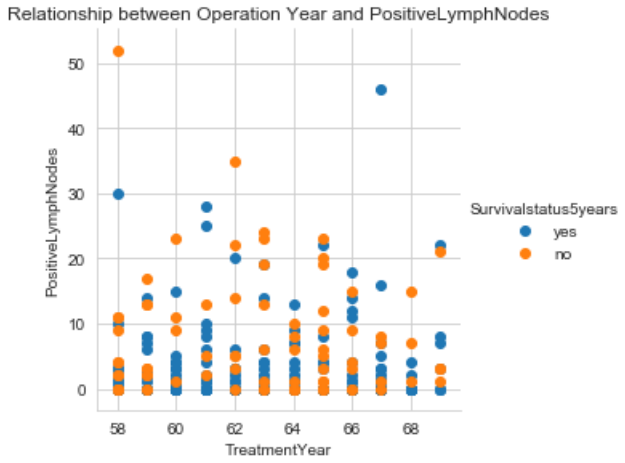
```
#2-D Scatter plot with color-coding for survival status class between Patient's Age and Year of Operati
on
sns.set_style("whitegrid")
sns.FacetGrid(df, hue='Survivalstatus5years',height=4) \
   .map(plt.scatter,"Patient Age","TreatmentYear") \
   .add_legend();
plt.title("Relationship between Patient's Age and Operation Year")
plt.show();
```



Observation:From the above plot we can see that there are too many overlapping points between Patient's age and Year of Operation Hence its very difficult to choose the relevant variable in this case

In [103]:

```
#2-D Scatter plot with color-coding for survival status class between Year of Operation  and Positive L
ymph Node
sns.set_style("whitegrid")
sns.FacetGrid(df, hue='Survivalstatus5years',height=4) \
   .map(plt.scatter,"TreatmentYear","PositiveLymphNodes") \
   .add_legend();
plt.title("Relationship between Operation Year and PositiveLymphNodes")
plt.show();
```



Observation:This plot looks a bit better but still there are a lot of overlapping points so cant consider any variables for classification

Pair Plot

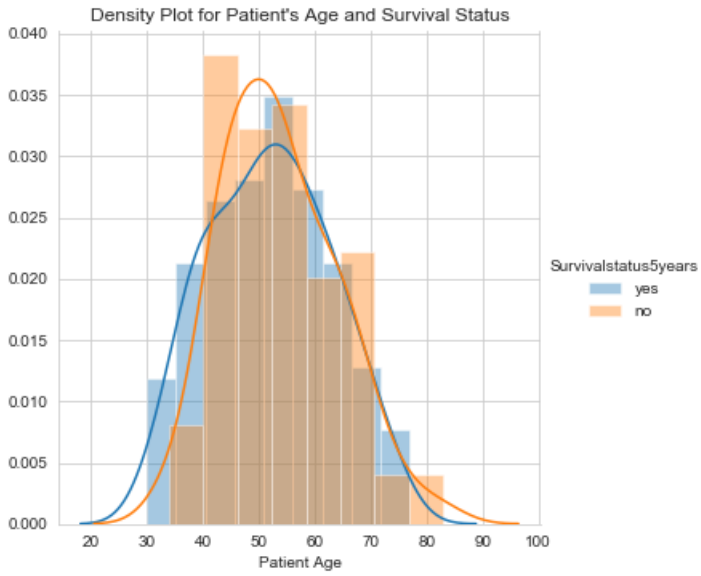
In [75]:

```
plt.close();
sns.set_style("whitegrid");
sns.pairplot(df, hue="Survivalstatus5years",height=3);
plt.show()
```

Observation: 1)From the above pair plot variables "Patient Age" and "Treatment Year" cannot be used for classifcation due to lots of overlapping points between them 2) PositiveLymphNodes looks like an appropriate variable for classification out of the rest as we can see that the survival rate for patients with more than 10 nodes can be separated from patients having less than 10 nodes

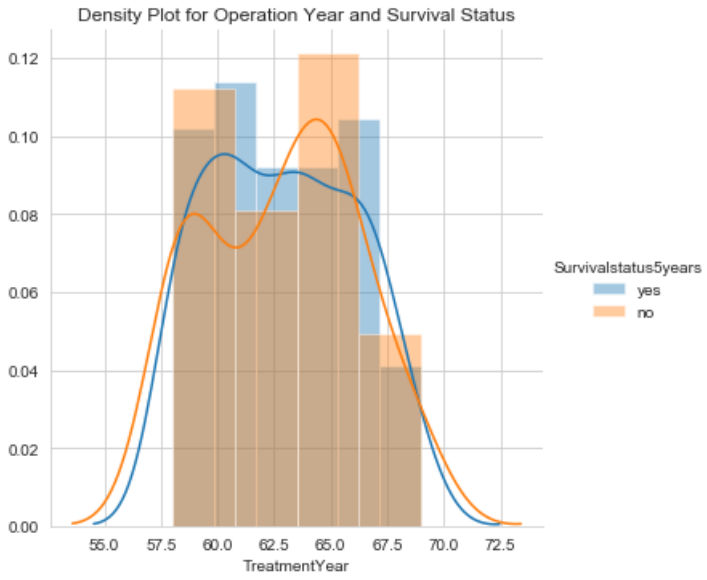
In [105]:

```
sns.FacetGrid(df,hue = "Survivalstatus5years", height=5) \
    .map(sns.distplot, "Patient Age") \
    .add_legend();
plt.title("Density Plot for Patient's Age and Survival Status")
plt.show();
```



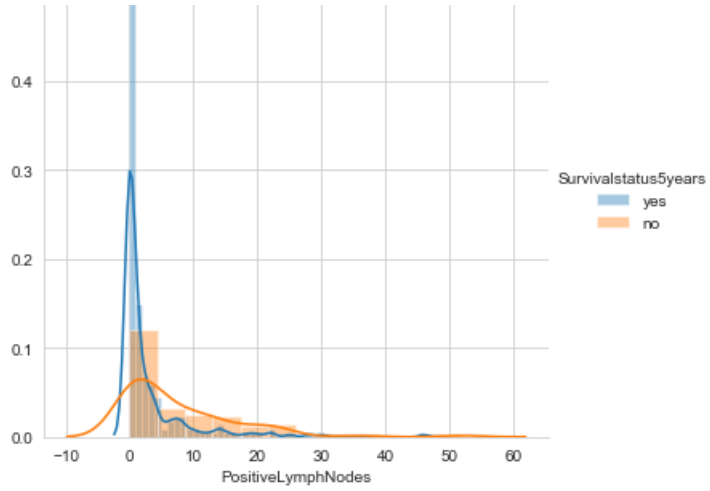
In [106]:

```
sns.FacetGrid(df,hue = "Survivalstatus5years", height=5) \
    .map(sns.distplot, "TreatmentYear") \
    .add_legend();
plt.title("Density Plot for Operation Year and Survival Status")
plt.show();
```



In [107]:

```
sns.FacetGrid(df,hue = "Survivalstatus5years", height=5) \
    .map(sns.distplot, "PositiveLymphNodes") \
    .add_legend();
plt.title("Density Plot for PositiveLymphNodes and SurvivalStatus")
plt.show();
```



Observations:From the above PDF Plots the PositiveLymphNodes is the most clear one as it gives the below insight. 1) The survival rate is higher than the non survival rate 2) The survival rate decreases if the PositiveLymphNodes exceeds more than 3

PDF & CDF Analysis

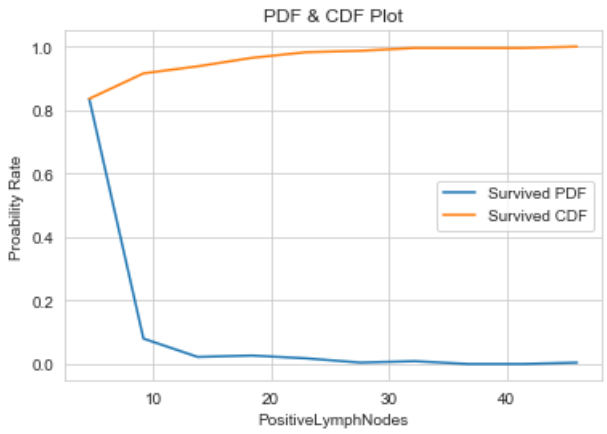
In [111]:

```
df_survival = df.loc[df["Survivalstatus5years"] == 'yes']
counts, bin_edges = np.histogram(df_survival['PositiveLymphNodes'], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
print(pdf)
print(bin_edges)

#Compute CDF
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)
plt.legend("Survivalstatus5years")
plt.legend(['Survived PDF','Survived CDF'])
plt.xlabel("PositiveLymphNodes")
plt.ylabel("Proability Rate")
plt.title("PDF & CDF Plot")

plt.show();
```

```
[0.83555556 0.08       0.02222222 0.02666667 0.01777778 0.00444444
 0.00888889 0.         0.         0.00444444]
[ 0.   4.6  9.2 13.8 18.4 23.  27.6 32.2 36.8 41.4 46. ]
```



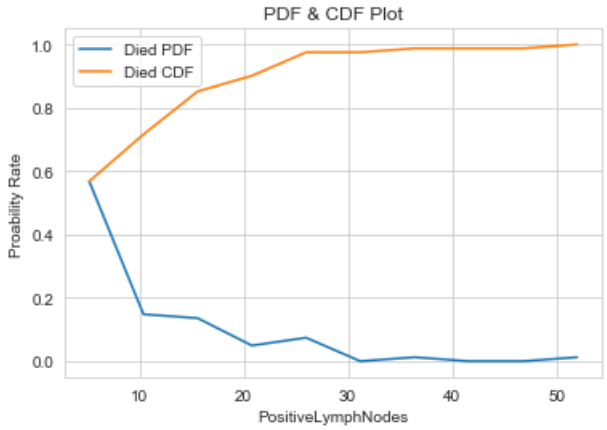
In [112]:

```
df_survival = df.loc[df["Survivalstatus5years"] == 'no']
counts, bin_edges = np.histogram(df_survival['PositiveLymphNodes'], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
```

```
print(pdf)
print(bin_edges)

#Compute CDF
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)
plt.legend("Survivalstatus5years")
plt.legend(['Died PDF','Died CDF'])
plt.xlabel("PositiveLymphNodes")
plt.ylabel("Proability Rate")
plt.title("PDF & CDF Plot")
plt.show();
```

```
[0.56790123 0.14814815 0.13580247 0.04938272 0.07407407 0.
 0.01234568 0.         0.         0.01234568]
[ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
```

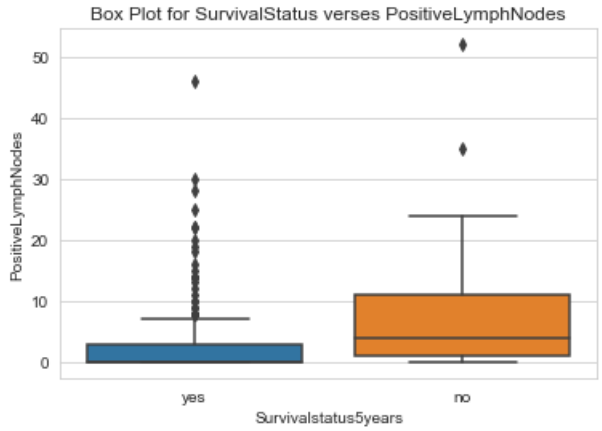


Observations:From the above graph we can say that patients having less than 10 Positive lymph nodes have 80% chances of survival rate than patients having more than 40 Positivelymphnodes

Box Plot

In [108]:

```
sns.boxplot(x='Survivalstatus5years',y='PositiveLymphNodes',data=df)
plt.title("Box Plot for SurvivalStatus verses PositiveLymphNodes")
plt.show()
```

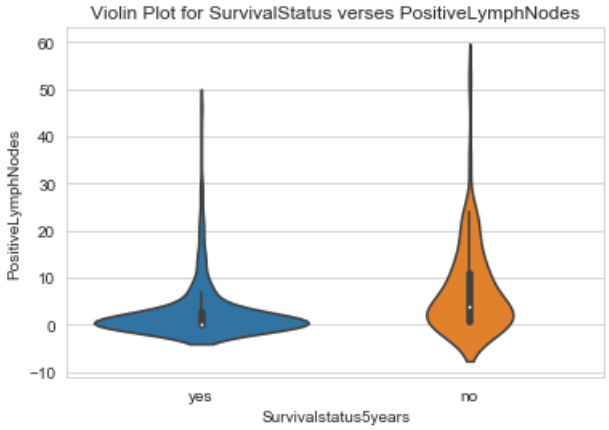


Observations:1) From the first box plot we can say that 75th percentile value for the PositiveLymphNodes is like 2 and less than 5, for which the survival rate is quite high. 2) From the second box plot we can say that 75th percentile value for the PositiveLymphNodes is 3 and higher for the Patient having less chances of survival

Violin Plot

In [109]:

```
sns.violinplot(x='Survivalstatus5years',y="PositiveLymphNodes",data=df,height=8)
plt.title("Violin Plot for SurvivalStatus verses PositiveLymphNodes")
plt.show()
```



Observations:1)From the first violin plot we can say that 50th percentile value for the PositiveLymphNodes is like 0 for which the survival rate is quite high than 75th percentile of the Patients having less than 3 PositiveLymphNodes 2)From the second violin plot we can say that 25th percentile value for the PositiveLymphNodes is like 1 for a Patient having less chances of survival,50th percentile value for patients likely to die with PositiveLymphNodes being below 4 and 75th percentile value for patients likely to die with PositiveLymphNodes being above 11

In [ ]:

```
```