Assignment Instructions:

1. Download Haberman Cancer Survival dataset from Kaggle. You may have to create a Kaggle account to donwload data. (https://www.kaggle.com/gilsousa/habermans-survival-data-set) or you can also run the below cell and load the data directly.
2. Perform a similar anlaysis as done in the reference notebook on this dataset.

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

y = survival_status

In [1]:

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import warnings
warnings.filterwarnings('ignore')
#%matplotlib inline
# we can read the data directly from raw github link
# we are also defining the name of the columns.
#make sure that your csv file and ipynb notebook are in the same folder. If t
df=pd.read_csv('haberman.csv',names=["age","operation_Year","axil_nodes","sur
df.head()
```

Out[1]:

	age	operation_Year	axil_nodes	survival_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

1.1 Analyze high level statistics of the dataset: number of points, numer of features, number of classes, data-points per class.

- You have to write all of your observations in Markdown cell with proper formatting.You can go through the following blog to understand formatting in markdown cells - https://www.markdownguide.org/basic-syntax/
- Do not write your observations as comments in code cells.
- Write comments in your code cells in order to explain the code that you are writing. Proper use of commenting can make code maintenance much easier, as well as helping make finding bugs faster.
- You can add extra cells using **Insert cell below command** in Insert tab. You can also use the shortcut Alt+Enter
- It is a good programming practise to define all the libraries that you would be using in a single cell

No of Data Points & Features

In [2]:

```
shape = df.shape
print('Total number of Points : ', shape[0])
print('Total number of Features : ', shape[1])
```

```
Total number of Points :  306
Total number of Features :  4
```

No. of Classes

In [3]:

```
predicted_class = df['survival_status'].unique()
print('Predictable Class values ', predicted_class)
print('Total number of unique classes ', len(predicted_class))
```

```
Predictable Class values  [1 2]
Total number of unique classes  2
```

No. of data points per classes

In [4]:

```
df['survival_status'].value_counts()
```

Out[4]:

```
1    225
2     81
Name: survival_status, dtype: int64
```

> From Above figures it can be inferred that total number of people survived are more than double of non-survived

1.2 - Explain the objective of the problem.

The aim of analysis is to find out the relation or pattern that may delineates some hindsights which can be used for future patients or other field

In [5]:

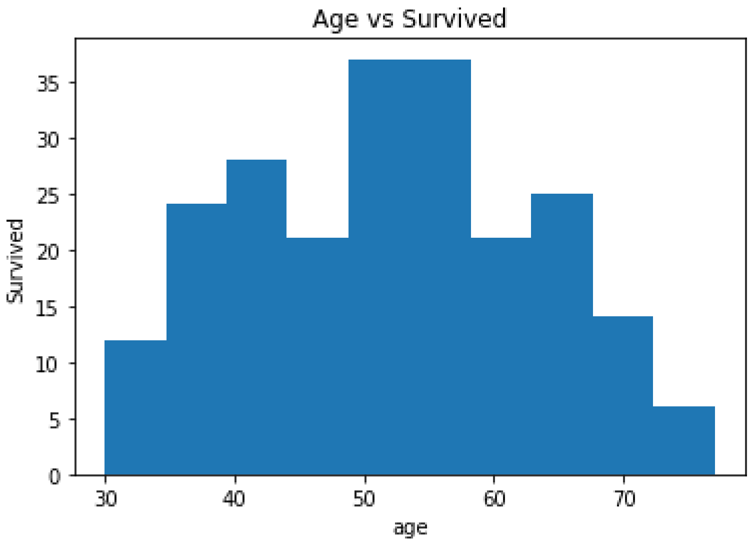
```
lo = df['age'].min()
hi = df['age'].max()
lo, hi
```

Out[5]:

```
(30, 83)
```

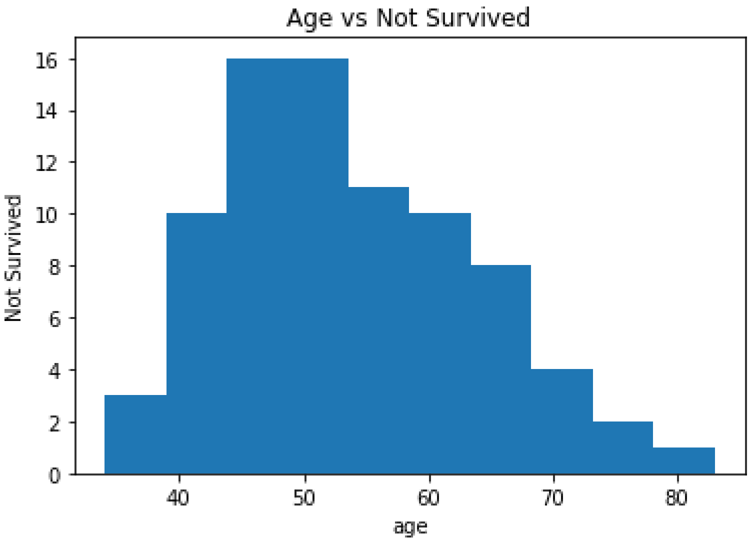
In [6]:

```
survived_age = df.loc[df['survival_status']==1]['age']

survived_age.plot(kind='hist', title="Age vs Survived")

plt.xlabel('age')
plt.ylabel('Survived')

plt.show()
```



In [7]:

```
survived_age = df[df['survival_status']==2]['age']

survived_age.plot(kind='hist', title='Age vs Not Survived')
plt.xlabel('age')
plt.ylabel('Not Survived')
plt.show()
```



Observations

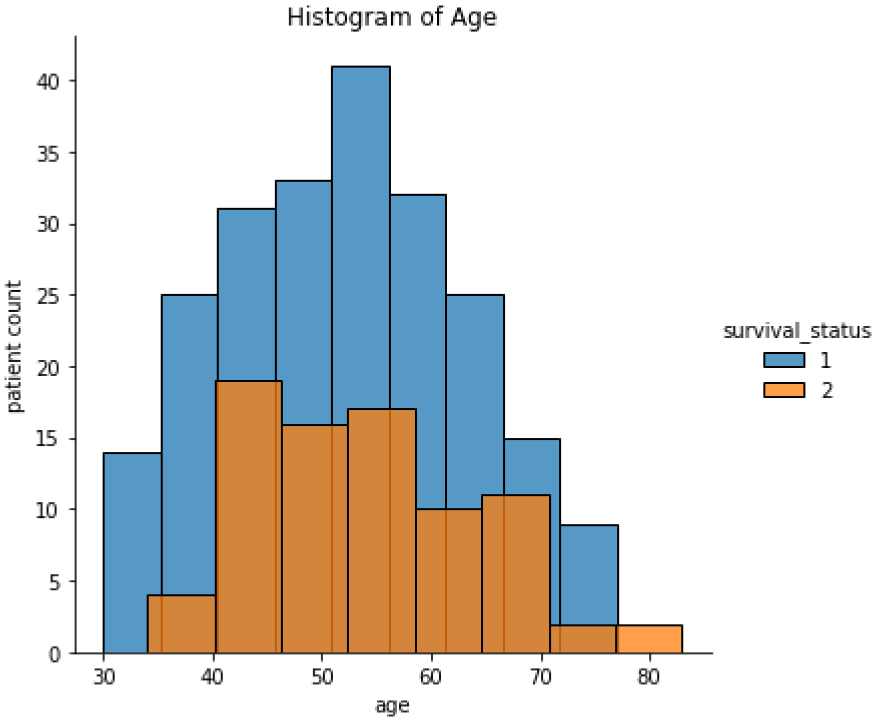
1. Age vs Survived graph seems more symmetric compared to Age vs Non-Survived
2. Age vs Non-Survived is more skewed towards Right Side
3. People in age group 50-60 are the one who survived maximally
4. People who did not survived are mostly in group of 40 to 60 years old

1.3 Perform Univariate analysis - Plot PDF, CDF, Boxplot, Voilin plots

- Plot the required charts to understand which feature are important for classification.
- Make sure that you add titles, legends and labels for each and every plots.
- Suppress the warnings you get in python, in that way it makes your notebook more presentable.
- Do write observations/inference for each plot.

In [8]:

```
sns.FacetGrid(df, hue="survival_status", size=5) \
    .map(sns.histplot, "age") \
    .add_legend()
plt.title('Histogram of Age')
plt.ylabel('patient count')
plt.show()
```

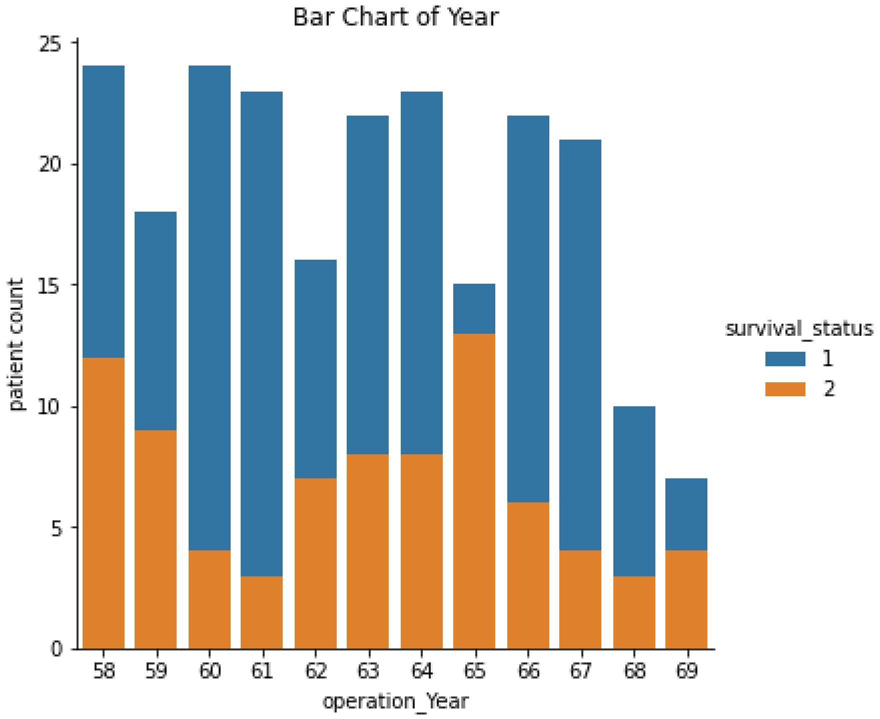


Observations

1. Major Category of Patient are 40-60 years old. Thus Middle-Age people are mostly suffered by the Breast Cancer.
2. Youngsters had high rate of success for surgery Whereas People in their Octogenerian could not survive the Surgery

In [19]:

```
sns.FacetGrid(df, hue="survival_status", size=5) \
    .map(sns.countplot, "operation_Year") \
    .add_legend()
plt.title('Bar Chart of Year')
plt.ylabel('patient count')
plt.show()
```

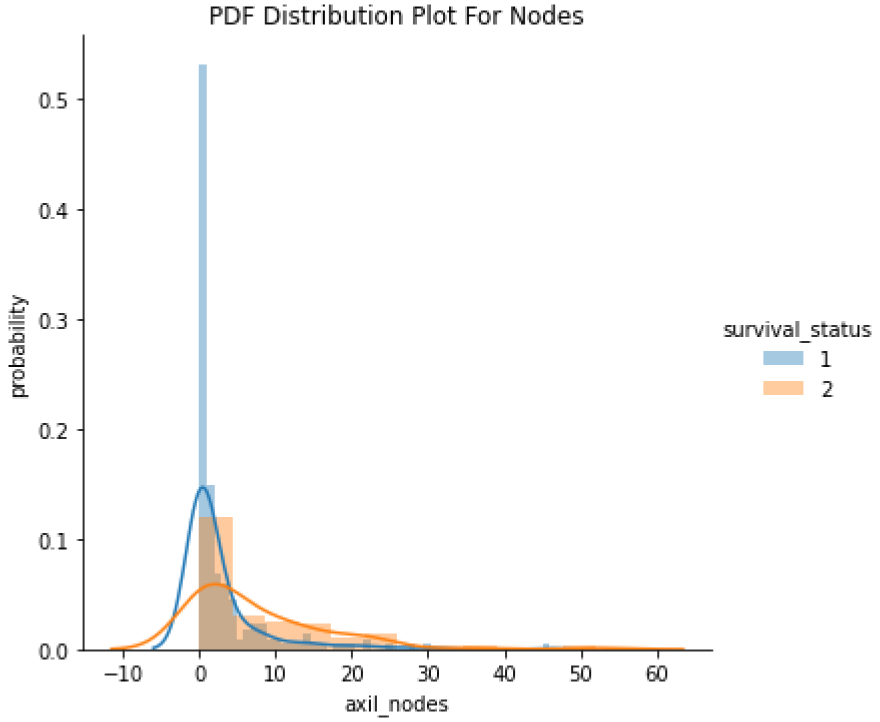


Observations

1. Most of patient died in year 65.
2. In the year 58, 59 & 69 almost 1 per 3 patient were insuccessful with their surgery
3. whereas in the year 60, 61, 66 & 67 majority of the patient were survived

In [20]:

```
sns.FacetGrid(df, hue="survival_status", size=5) \
    .map(sns.distplot, "axil_nodes") \
    .add_legend()
plt.title('PDF Distribution Plot For Nodes')
plt.ylabel('probability')
plt.show()
```

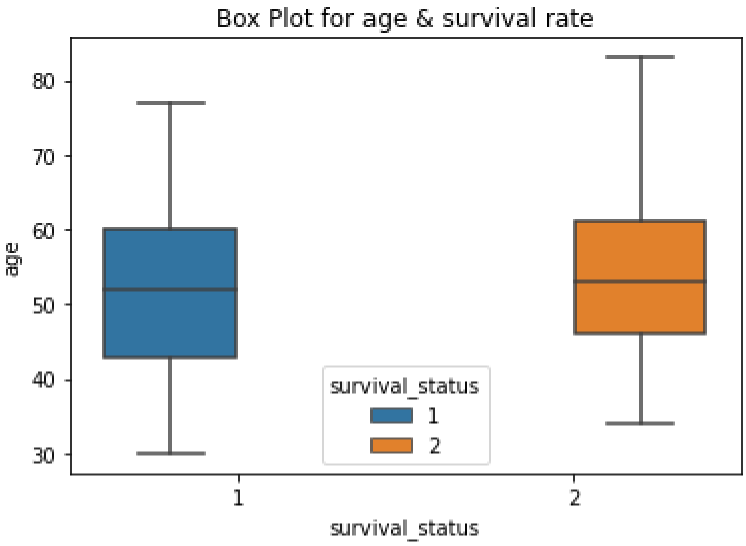


Observations

1. Approximately 50% of Pateints with 0 axil_nodes were having the success rate for Surgery
2. Maximum operation failure was encountered for patients with axil nodes count in between 0 to 20
3. As axile nodes count increases survival rate reduces & failure chances increases

In [31]:

```
sns.boxplot(x="survival_status",y='age', data=df, hue='survival_status')
plt.title('Box Plot for age & survival rate')
plt.show()
```

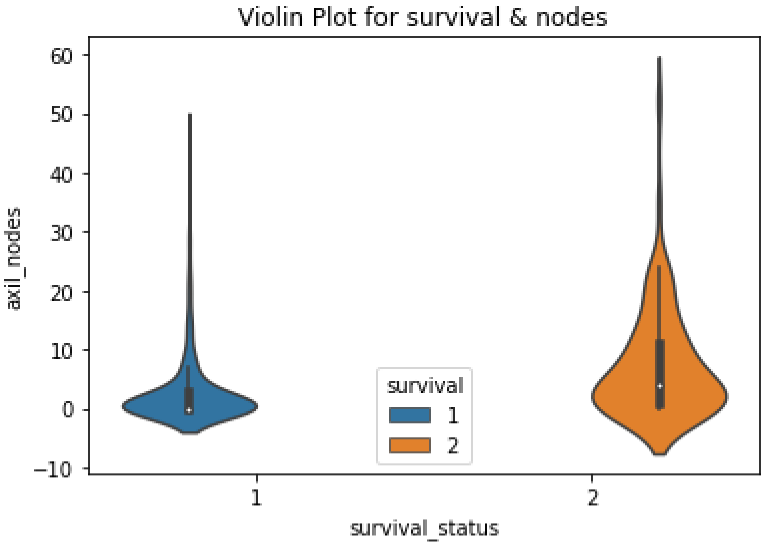


Observations

1. Major of Survived & Non-Survived, Patient lies between 45-60 Years.
2. Thus indicating that brest-cancer being diaognised positive for People falling in age 45 to 60 is High.

In [37]:

```
sns.violinplot(x="survival_status", y="axil_nodes", data=df, size=8, hue="sur
plt.title('Violin Plot for survival & nodes')
plt.legend(loc='lower center', title="survival")
plt.show()
```

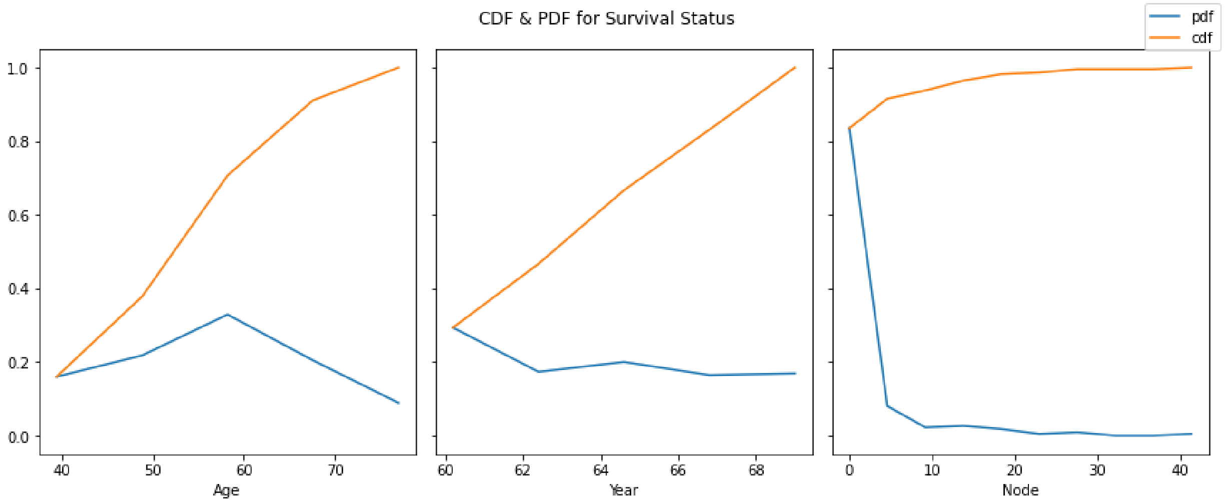


Observations

1. As axil nodes increases from 0 towards 60 chances for surgery failure are also increases mildly

In [38]:

```
status_survived = df.loc[df['survival_status']==1]
status_death = df.loc[df['survival_status']==2]

fig, axes = plt.subplots(1, 3, sharey=True, figsize=(12,5))

fig.suptitle('CDF & PDF for Survival Status')

# Survival Age PDF & CDF
counts, bin_edges = np.histogram(status_survived['age'], bins=5)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
axes[0].plot(bin_edges[1:],pdf)
axes[0].plot(bin_edges[1:], cdf)
axes[0].set_xlabel('Age')

# Survival Operation Year PDF & CDF
counts, bin_edges = np.histogram(status_survived['operation_Year'], bins=5)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
axes[1].plot(bin_edges[1:],pdf)
axes[1].plot(bin_edges[1:], cdf)
axes[1].set_xlabel('Year')

# Survival Axil Nodes PDF & CDF
counts, bin_edges = np.histogram(status_survived['axil_nodes'], bins=10)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
axes[2].plot(bin_edges[0:-1],pdf, label='pdf')
axes[2].plot(bin_edges[0:-1],cdf, label='cdf')
axes[2].set_xlabel('Node')

fig.tight_layout()
handles, labels = axes[-1].get_legend_handles_labels()
fig.legend(handles, labels, loc='upper right')
plt.show()
```



Observations

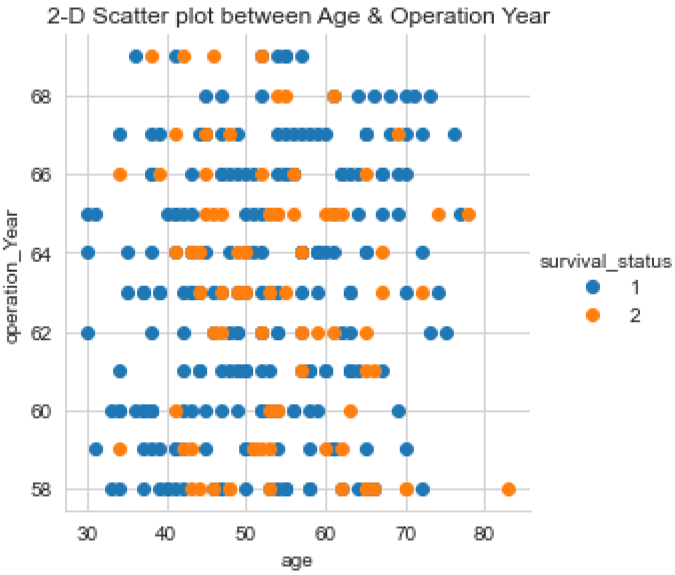
1. About 80% of Population who survived, were below age of 65
2. 80% of Survivals were having axil node count of 0

1.4 Perform Bivariate analysis - Plot 2D Scatter plots and Pair plots

- Plot the required Scatter plots and Pair plots of different features to see which combination of features are useful for clasification task
- Make sure that you add titles, legends and labels for each and every plots.
- Suppress the warnings you get in python, in that way it makes your notebook more presentable.
- Do write observations/inference for each plot.

In [39]:

```
# 2-D Scatter plot
sns.set_style("whitegrid");
sns.FacetGrid(df, hue="survival_status", size=4) \
   .map(plt.scatter, "age", "operation_Year") \
   .add_legend();
plt.title('2-D Scatter plot between Age & Operation Year')
plt.show();
```

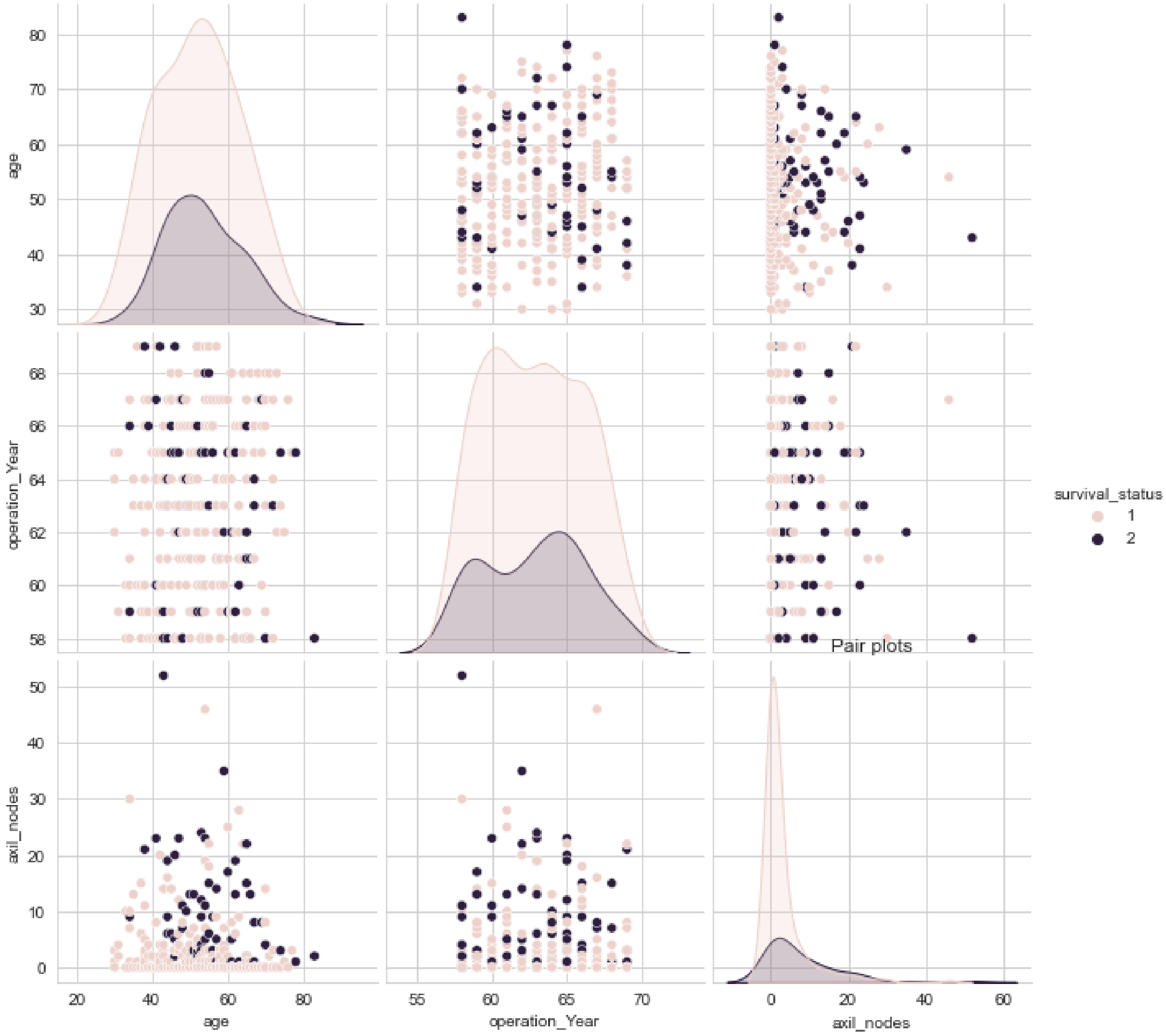


Observation

1. Operation In the year 66-68 were successful for Most of the Senior Citizens
2. Also in the period between 60 & 62 most of the operations were successful

In [40]:

```
plt.close();
sns.pairplot(df, hue="survival_status", size=3);
plt.title('Pair plots')
plt.show()
```



Observation

1. In the period between 60-62, around 95% of patient survived post operation
2. The number of people survived is apparantly more than number of people didn't

1.5 Summarize your final conclusions of the Exploration

- You can desrcibe the key features that are important for the Classification task.
- Try to quantify your results i.e. while writing observations include numbers,percentages, fractions etc.
- Write a brief of your exploratory analysis in 3-5 points
- Write your observations in english as crisply and unambigously as possible.

Conclusion (Observation)

1. All in all it can be said that Mostly middle aged person belonging to the age group of 40 - 60 were suffered by breast cancer.
2. Although there were Young Patients but most of them survived the operation.
3. Moreover It was observed that 80% of patient who had successful operation were below the age of 65 & also they were having axil counts of 0.
4. To add to this as axil count increases the chances of survival decreases moderately, altogether