About Dataset

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.

Objective of the problem

Main objective of this problem is to find whether the patients who have undergone breast Cancer surgery will be survived or not

We will be performing Exploratory Data Analysis(EDA) on the Data that we have obtained from Kaggle.

Link for Dataset: https://www.kaggle.com/datasets/gilsousa/habermans-survival-data-set

```
''' Importing required Libraries '''

import warnings
warnings.filterwarnings("ignore")

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

```
''' I have stored the Dataset in my local and loaded data into  below
DataFrame
```

```
''' Loading the data into a Pandas DataFrame(df)'''

df = pd.read_csv('haberman.csv',
names=["age","operation_year","axil_nodes","survival_status"]) #
defining names for columns using 'names'
df = df.iloc[1: , :] # selecting all the rows from row num = 1
(dropping first row in order to avoid default column names present in
our data)
df = df.astype(int) # converting into type of int
df.head()
```

```
   age  operation_year  axil_nodes  survival_status
1   30              64           1                1
2   30              62           3                1
3   30              65           0                1
4   31              59           2                1
5   31              65           4                1
```

```
df.info()
```

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

```
df.columns # list of columns present in our Data
```

```
Index(['age', 'operation_year', 'axil_nodes', 'survival_status'],
dtype='object')
```

About the columns in our data:

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) --> A positive axillary node is a lymph node in the area of the armpit (axilla) to which cancer has spread. This spread is determined by surgically removing some of the lymph nodes and examining them under a microscope to see whether cancer cells are present.
4. ***Survival_status***: Survival status (class attribute) 1 = the patient survived 5 years or longer 2 = the patient died within 5 year

```
df.shape # shape of our data
```

```
(306, 4)
```

Observation:

1. We have a total of 306 observations and 4 columns

```
''' checking whether there are any null values in our Data '''
df.isna().sum()
```

```
age                0
operation_year     0
axil_nodes         0
survival_status    0
dtype: int64
```

Observation:

1. We dont have any Null values in our Data in any of the columns

```
df.info() # printing some more info of our Data
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306 entries, 1 to 306
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype
```

```
---  ------           --------------  -----
 0   age              306 non-null    int64
 1   operation_year   306 non-null    int64
 2   axil_nodes       306 non-null    int64
 3   survival_status  306 non-null    int64
dtypes: int64(4)
memory usage: 9.7 KB
```

```
df.describe()
```

```
              age  operation_year  axil_nodes  survival_status
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
```

```
df.survival_status.unique() # we have 2 survival status
```

```
array([1, 2])
```

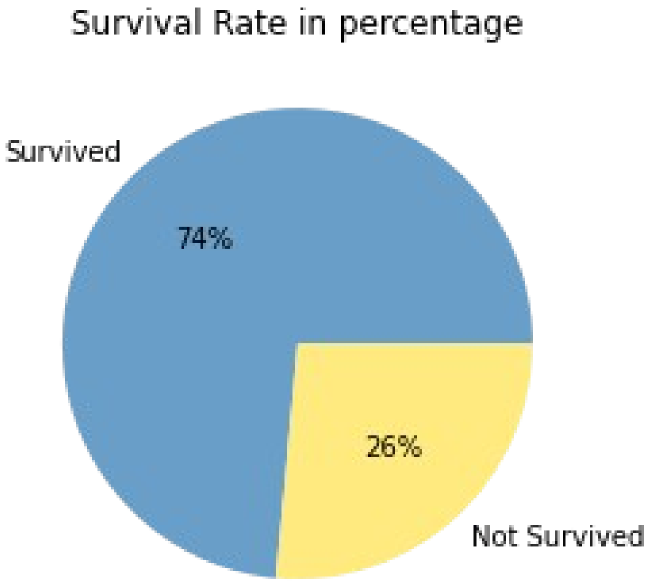
```
df.survival_status.value_counts()
```

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

Observation:

1. Out of 306 observations, 225 people survived for more than 5 years
2. And 81 people died within 5 years

```
slices = df["survival_status"].value_counts()

plt.pie(x=slices, labels=["Survived", "Not Survived"],
colors=["#699ec9", "#ffea80"], autopct="%1.0f%%")
plt.title("Survival Rate in percentage")


plt.show()
```



Observation:

1. From the Pie chart we can observe that 74% of patients of survived which is 48% more than Not Survived patients

```
# mean

print("Mean for all the 4 columns in our Data:")
print("="*40)

print('Age:', np.mean(df['age']))
print('Operation Year:', np.mean(df['operation_year']))
print('Axil Nodes:', np.mean(df["axil_nodes"]))
print('Survival Status:', np.mean(df["survival_status"]))
print('\n')
#=================================================================#

#standard deviations

print("Standard Deviations for all the 4 columns in our Data:")
print("="*55)

print('Age:', np.std(df['age']))
print('Operation Year:', np.std(df['operation_year']))
print('Axil Nodes:', np.std(df["axil_nodes"]))
print('Survival Status:', np.std(df["survival_status"]))
print('\n')
#=================================================================#

#median
```

```
print("Median for all the 4 columns in our Data:")
print("="*42)

print('Age:', np.median(df['age']))
print('Operation Year:', np.median(df['operation_year']))
print('Axil Nodes:', np.median(df["axil_nodes"]))
print('Survival Status:', np.median(df["survival_status"]))
```

```
Mean for all the 4 columns in our Data:
========================================
Age: 52.45751633986928
Operation Year: 62.85294117647059
Axil Nodes: 4.026143790849673
Survival Status: 1.2647058823529411


Standard Deviations for all the 4 columns in our Data:
=======================================================
Age: 10.785785203631832
Operation Year: 3.244090833563246
Axil Nodes: 7.177896092811143
Survival Status: 0.4411764705882353


Median for all the 4 columns in our Data:
==========================================
Age: 52.0
Operation Year: 63.0
Axil Nodes: 1.0
Survival Status: 1.0
```

Observation:

1. Average age of patients is 52

```
# quantiles

print("Quantiles for all the 4 columns in our Data(0%, 25%, 50%, 
100%):")
print("="*65)

print('Age:', np.percentile(df['age'],np.arange(0,100,25)))
print('Operation Year', 
np.percentile(df['operation_year'],np.arange(0,100,25)))
print('axil_nodes', 
np.percentile(df["axil_nodes"],np.arange(0,100,25)))
print('survival_status', 
np.percentile(df["survival_status"],np.arange(0,100,25)))
```

```
Quantiles for all the 4 columns in our Data(0%, 25%, 50%, 100%):
=================================================================
Age: [30.   44.   52.   60.75]
Operation Year [58.   60.   63.   65.75]
axil_nodes [0. 0. 1. 4.]
survival_status [1. 1. 1. 2.]
```

```
# 90th percentile
print("90th percentile value of all the 4 columns we have:")
print("="*50)

print('Age', np.percentile(df['age'],90))
print('Operation year', np.percentile(df['operation_year'],90))
print('Axil Node', np.percentile(df["axil_nodes"],90))
print('Survival Status', np.percentile(df["survival_status"],90))
```

```
90th percentile value of all the 4 columns we have:
==================================================
Age 67.0
Operation year 67.0
Axil Node 13.0
Survival Status 2.0
```

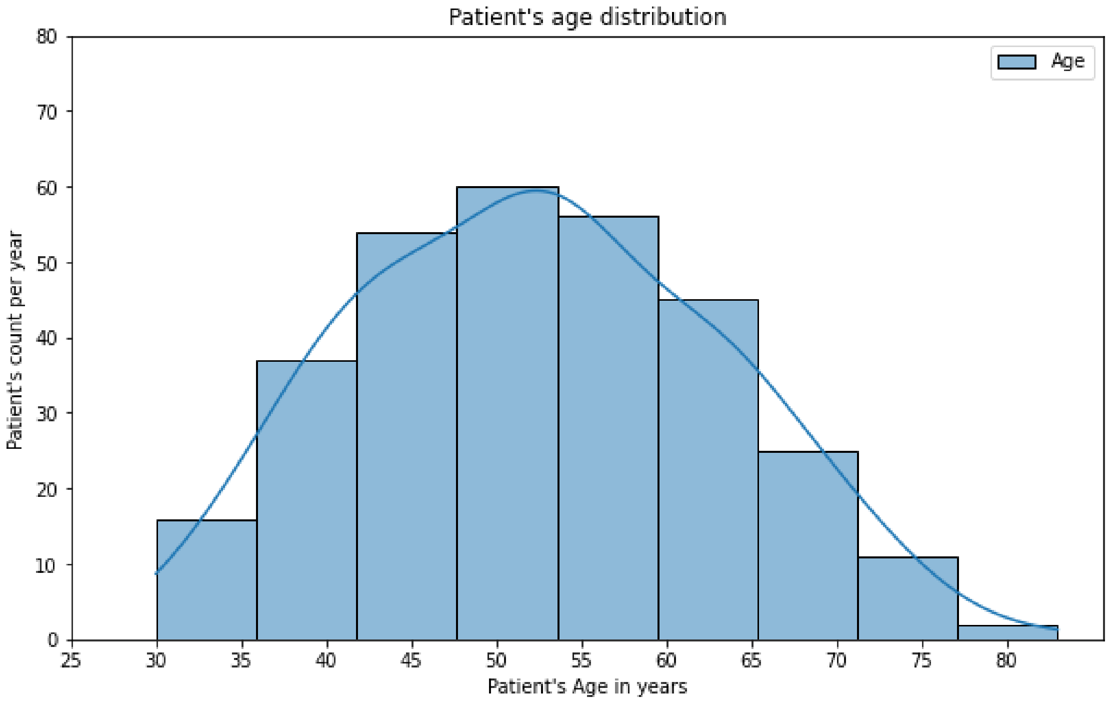
Observations:

1. 90% of patients age is around 67
2. 90% of patients has died within 5 years, as we have survial status 2 for 90% of the patients
3. 90% of patients has detected axil nodes = 13

Uni-variate Analysis

Patient Age

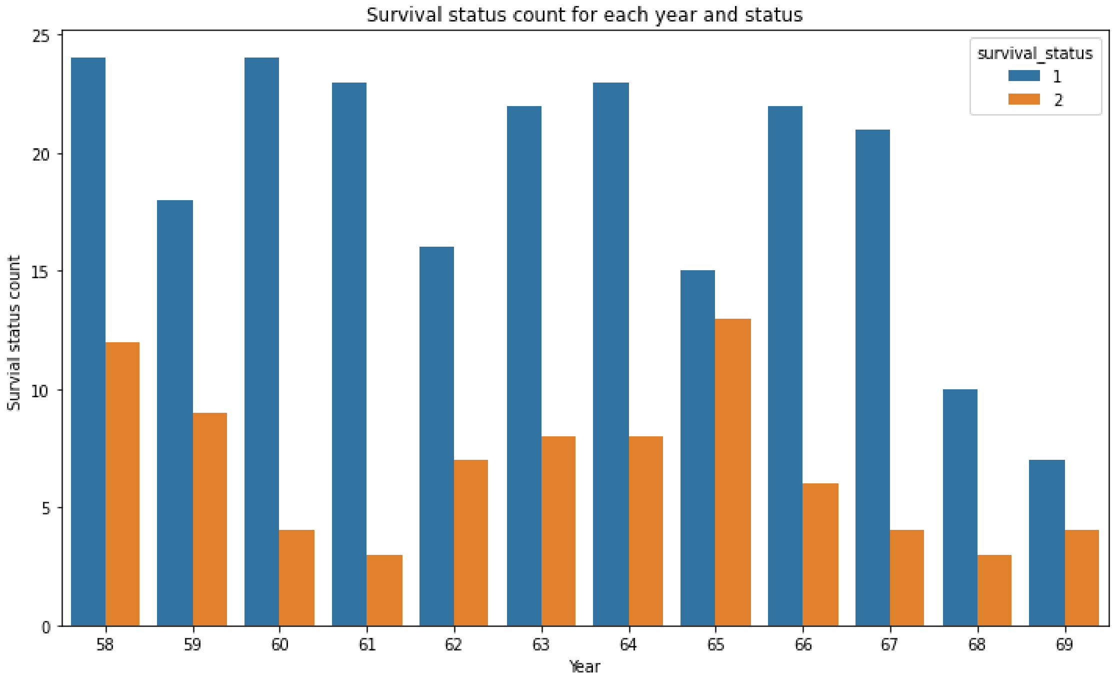
```
plt.figure(figsize=(10, 6))  # Figure size: width, height

sns.histplot(x = df["age"], label = "Age", bins = 9, kde = True)
plt.xlabel("Patient's Age in years")
plt.ylabel("Patient's count per year")
plt.title("Patient's age distribution")
plt.xticks(ticks=range(25, 85, 5))
plt.yticks(ticks=range(0, 90, 10))
plt.legend()


plt.show()
```



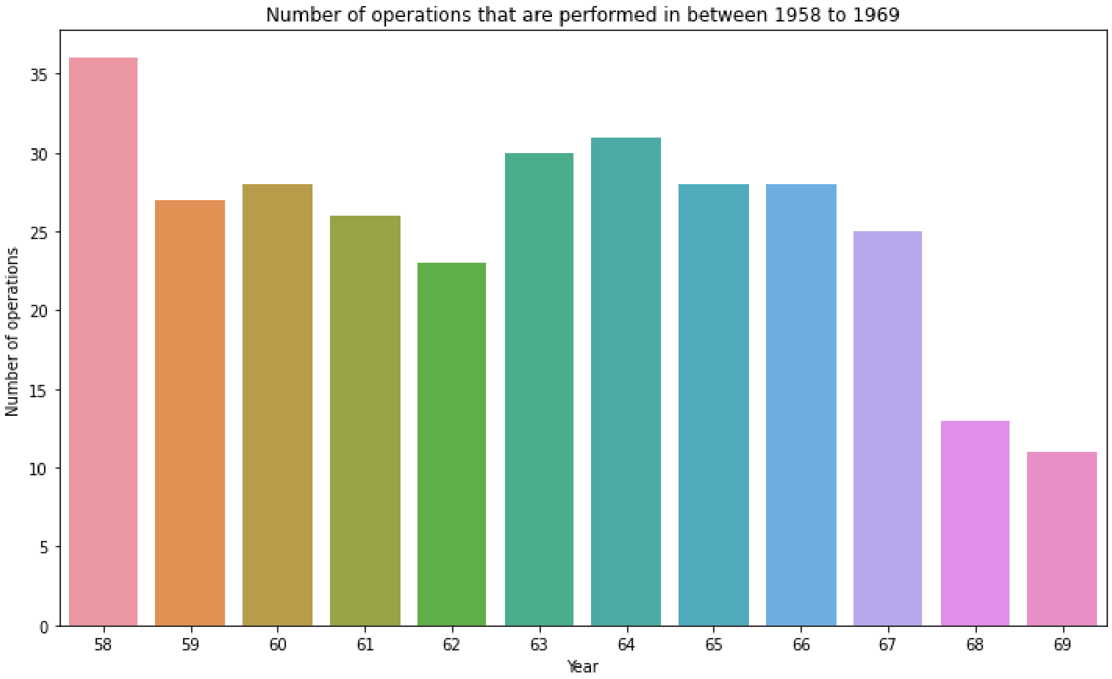
Observation:

1. From this we can understand that most of the patients who have undergone operation are in between 47 to 60 years.
2. We have a total of nearly 60 patients who are having age of 50 years

```
plt.figure(figsize=(12, 7))
sns.countplot(data=df, x='operation_year', hue='survival_status')

plt.xlabel("Year")
plt.ylabel("Survial status count ")
plt.title("Survival status count for each year and status")
plt.show()
```



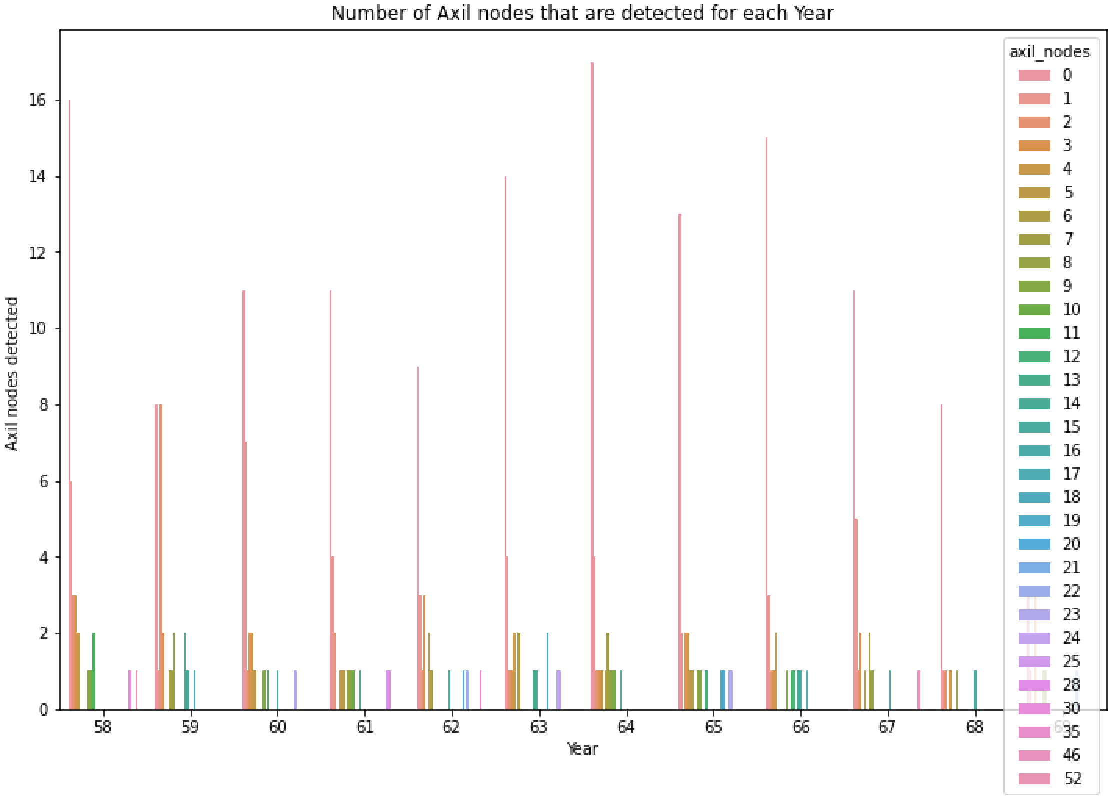
```
#df.groupby(['operation_year'])['axil_nodes'].count() # grouping
number of operation

plt.figure(figsize=(12, 7))
sns.countplot(data = df, x='operation_year')

plt.xlabel("Year")
plt.ylabel("Number of operations ")
plt.title("Number of operations that are performed in between 1958 to
1969")
plt.show()
```



Observation:

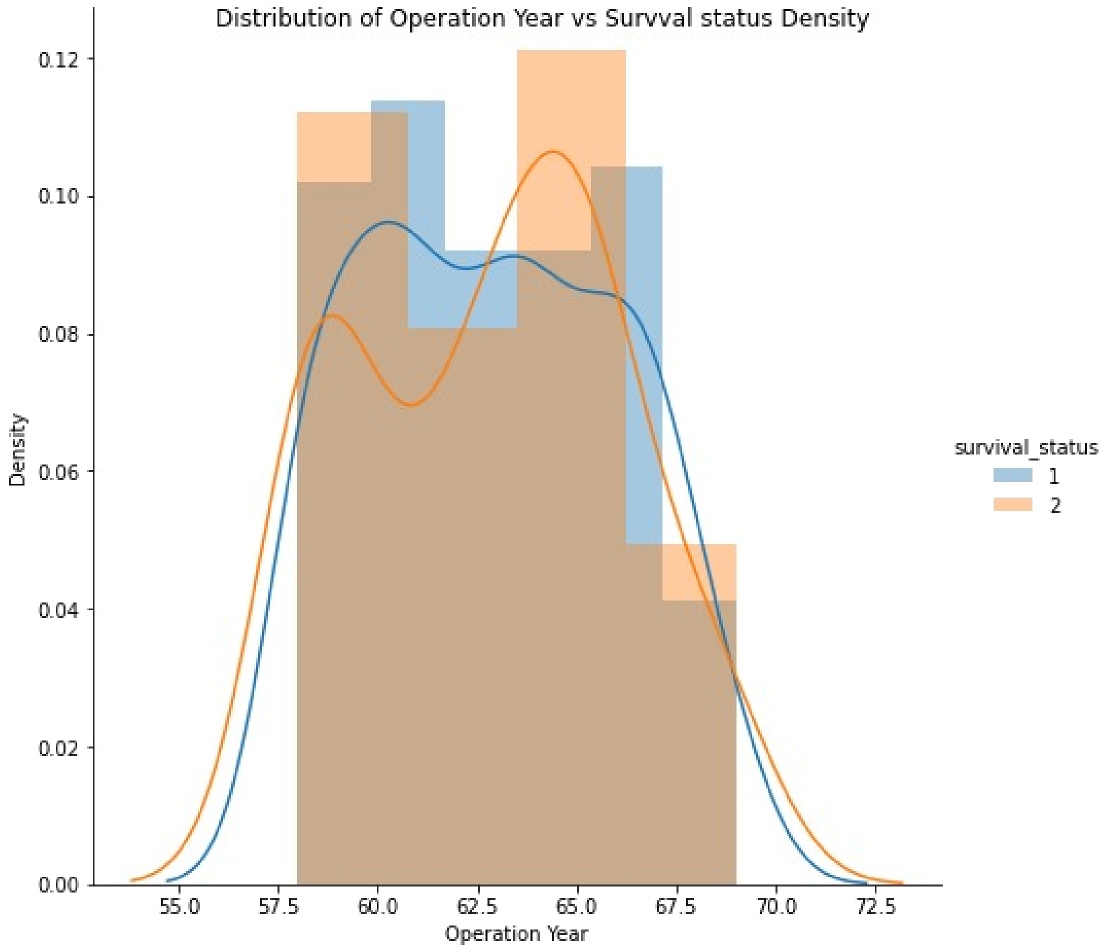
1. From this we can understand that number of operations that are performed in the year 1969 are very much lesser than in the year 1958, this may be due to the fact that there was more awarness about this disease and patients got cured even without an operation
2. In the year 1958, 36 number of patients have undergone surgery which is highest among all the other years

```
plt.figure(figsize=(12, 8))
sns.countplot(data=df, x='operation_year', hue='axil_nodes')

plt.xlabel("Year")
plt.ylabel("Axil nodes detected ")
plt.title("Number of Axil nodes that are detected for each Year")
plt.show()
```



Patient, Operation_year Distribution

```
fg = sns.FacetGrid(df,hue="survival_status",size = 7)\
     .map(sns.distplot,"operation_year")\
     .add_legend()

fg.fig.suptitle('Distribution of Operation Year vs Survval status
Density') # adding title
plt.xlabel("Operation Year")
plt.show();
```

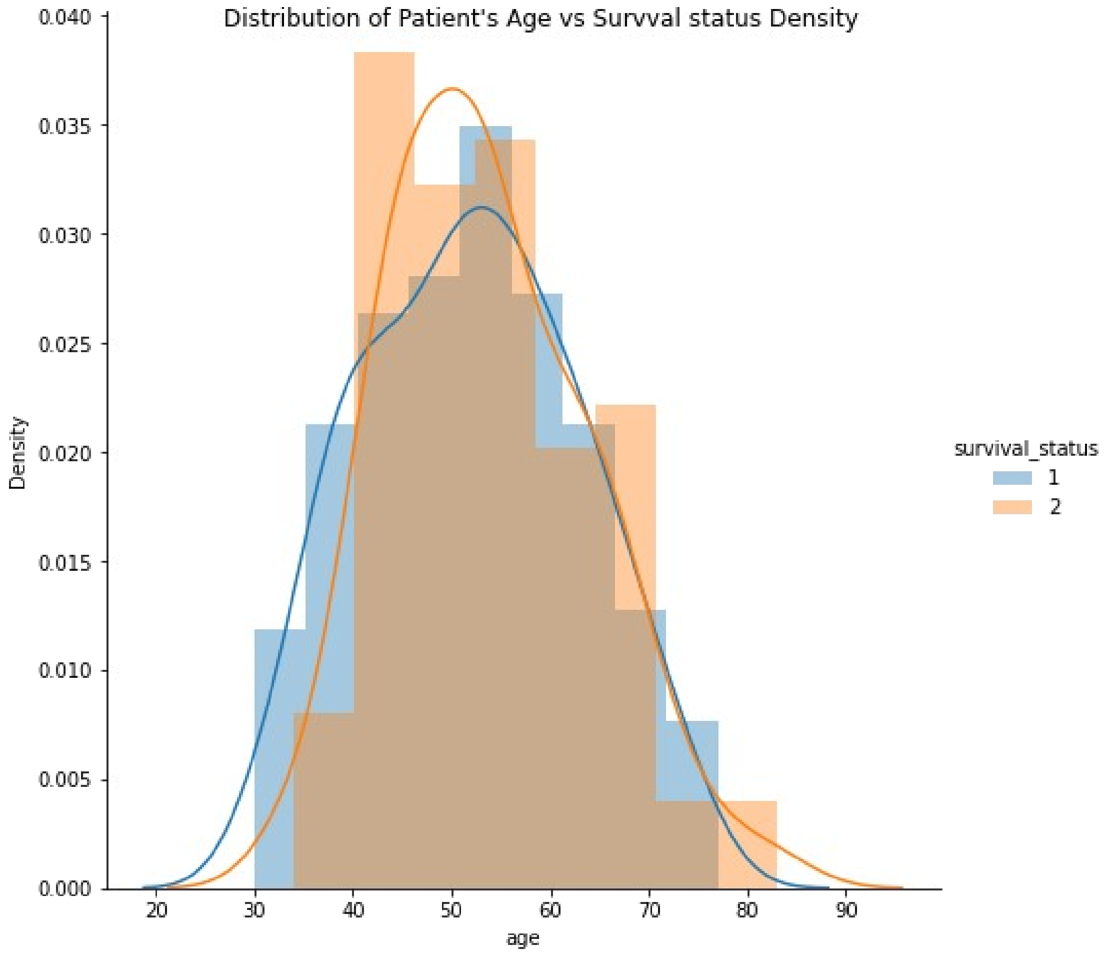


Observation:

1. We cant get much conclusion on the survial status of the patients based on their Operation year as most of the values are overlapped

Patient, Age distribution

```
fg = sns.FacetGrid(df,hue="survival_status",size = 7 )\
     .map(sns.distplot,"age")\
     .add_legend()

fg.fig.suptitle("Distribution of Patient's Age vs Survval status
Density") # adding title
plt.show();
```

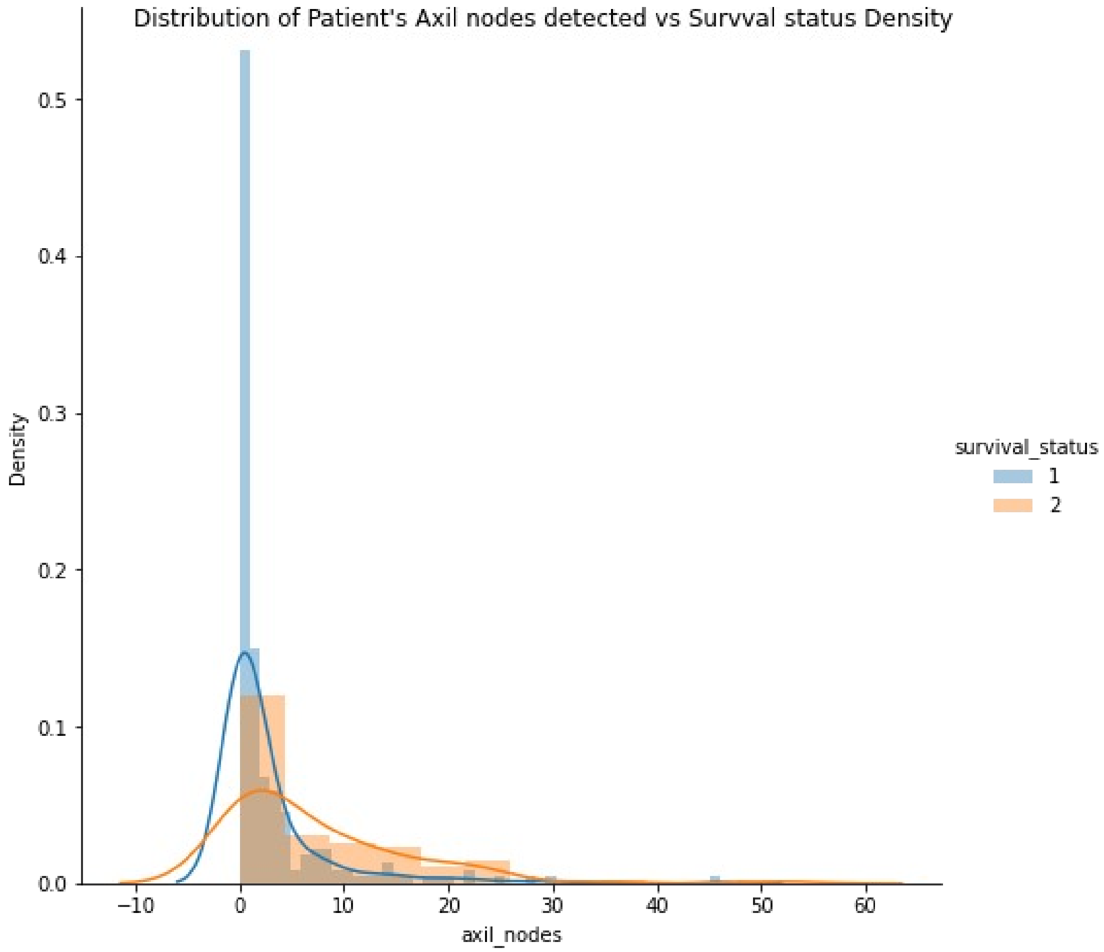


Observation:

1. We can observe that more patinets age in the range of 40 to 45 are having survival status = 2, i.e, died within 5 years of operation

Number of nodes Distribution

```
fg = sns.FacetGrid(df,hue = "survival_status",size = 7 )\
     .map(sns.distplot,"axil_nodes")\
     .add_legend()

fg.fig.suptitle("Distribution of Patient's Axil nodes detected vs
Survval status Density") # adding title
plt.show();
```

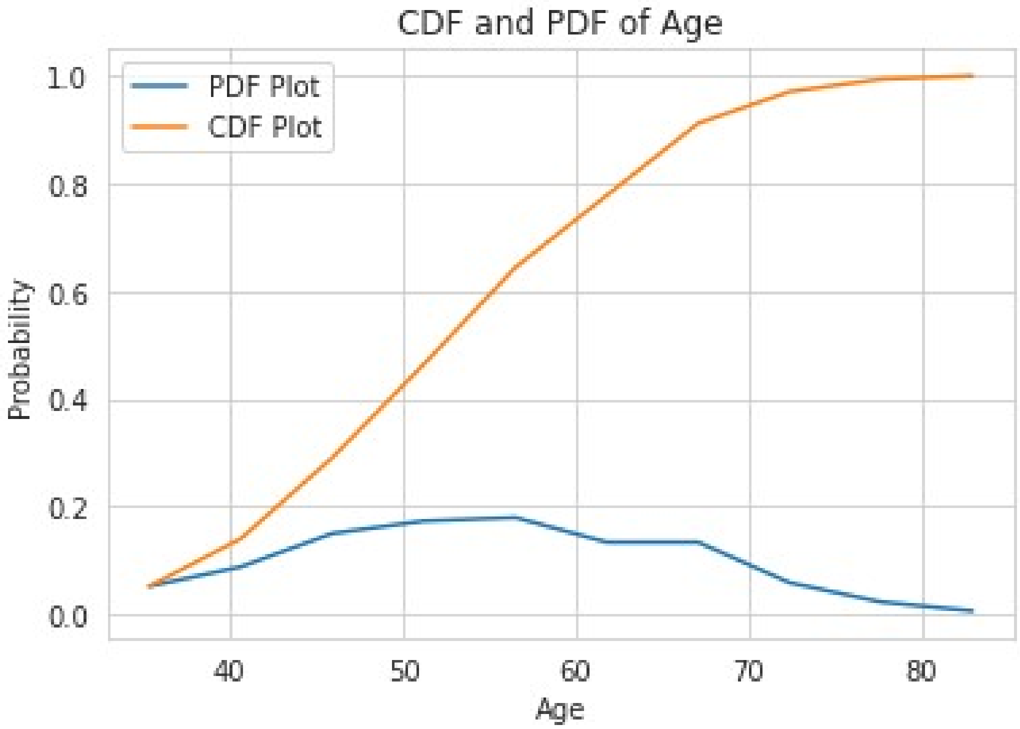


Observations:

1. We can observe that patients who are having around 0 number of **axil nodes** have very much higher survival rate(have more number of survival ststus =1)
2. From this we can tell that this is a very important feature to determine Patients survial rate

Patient Age CDF and PDF

```
# compute pdf

counts, bin_edges = np.histogram(df["age"],bins=10)
pdf  = counts/sum(counts)
#print(pdf)
#print(bin_edges)

#compute cdf
cdf = np.cumsum(pdf)
#print(cdf)

#plotting pdf nd cdf
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)

plt.title("CDF and PDF of Age")
plt.gca().legend(('PDF Plot','CDF Plot'))
plt.xlabel("Age")
plt.ylabel("Probability")
plt.show()
```

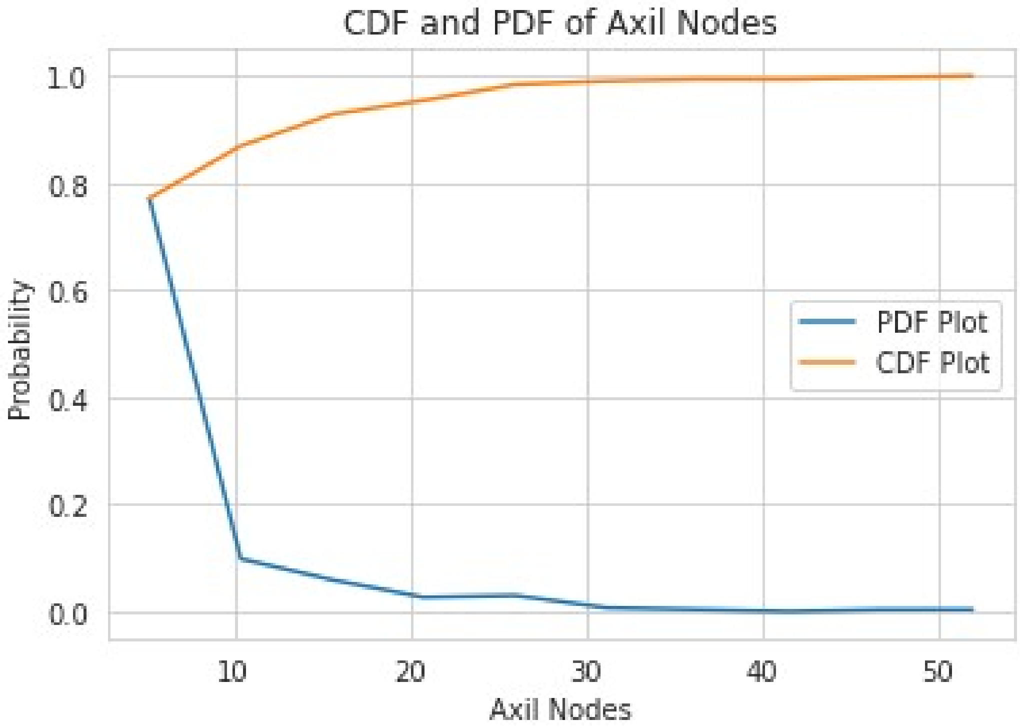


Observation:

1. Nearly 95% of the patients age is less than or equal to 70 and rest 5% of patients are having age greater than 70

Number of axil nodes CDF and PDF

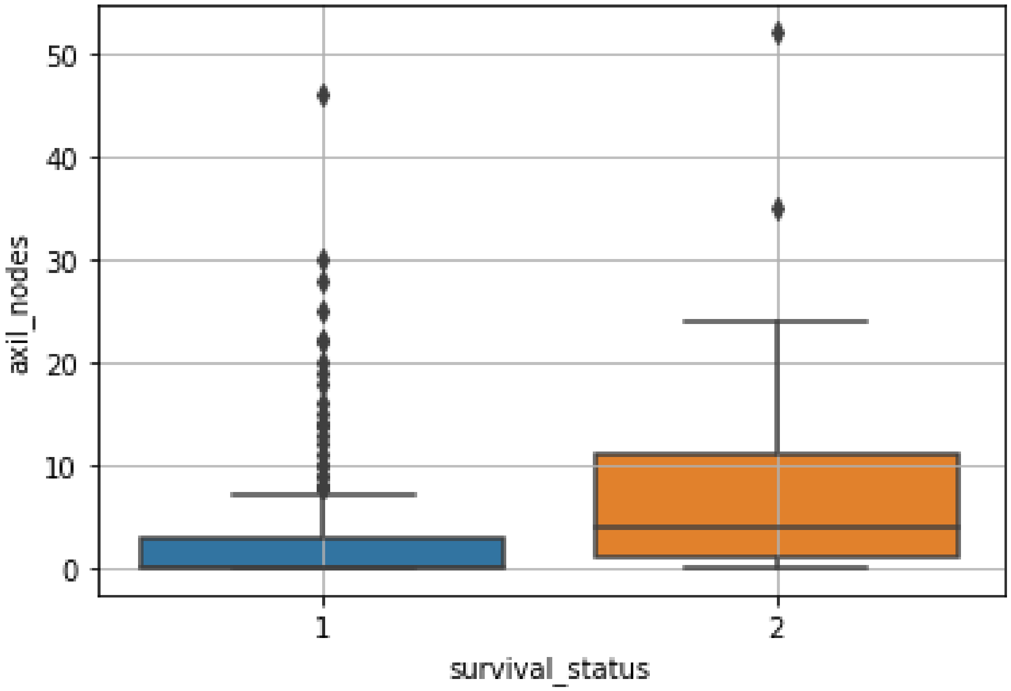
```
count,bin_edges = np.histogram(df["axil_nodes"],bins=10)

pdf = count/sum(counts)
cdf = np.cumsum(pdf)

plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)

plt.title("CDF and PDF of Axil Nodes")
plt.gca().legend(('PDF Plot','CDF Plot'))
plt.xlabel("Axil Nodes")
plt.ylabel("Probability")
plt.show()
```



Observations:

1. Most of the patients are having Axil nodes less than 10 in number
2. More than 82% of the patients are having Axil nodes greater than 10

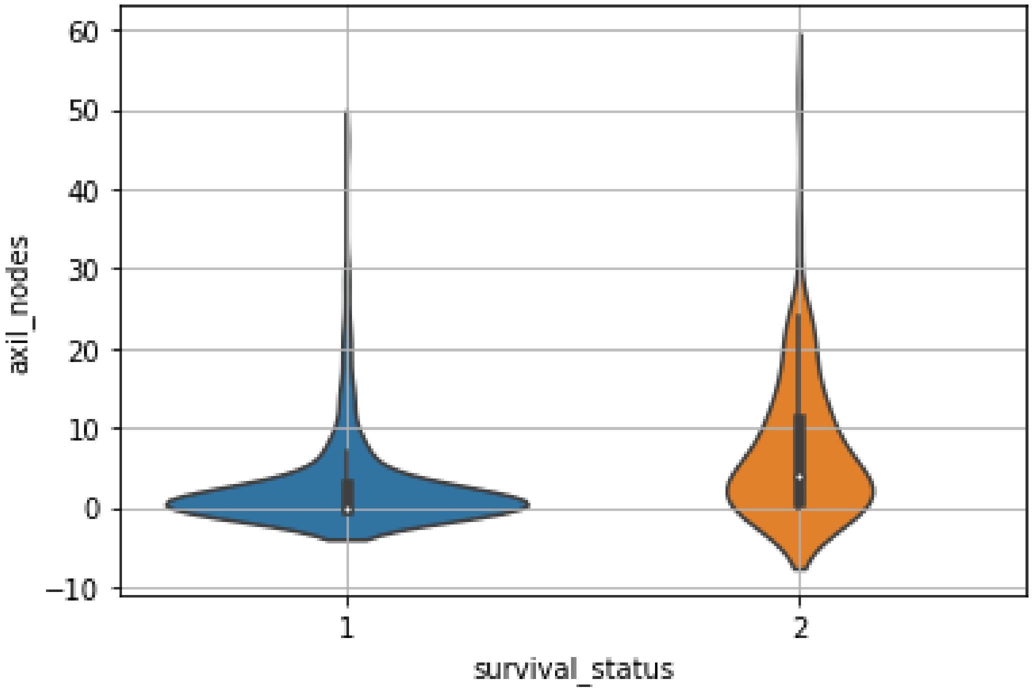
```
sns.boxplot(data = df, x="survival_status", y="axil_nodes")
plt.grid()
plt.show()
```



Observations:

1. We can observer that patients who are having axil nodes in the range of 0 to 5 are having higher survial rate(survial staus =1, i.e, survived for more than 5 years after the operation)
2. Patients who are having number of axil nodes greater than 5 didnot survive for more than 5 years
3. Nearly 75% of the patients who has survived are having axil nodes less than 5 in number

```
# a = df.loc[df['survival_status'] == 1]

''' violin plot is a combination of Box plot and Histogram, where we have Boxplot inside with 25, 50 and 75th percentile similar to boxplot and
    on the sides it will have PDF(Histograms) for those values in survial_status column'''

sns.violinplot(data = df, x ="survival_status", y = "axil_nodes",size=16)
plt.grid()
plt.show()
```



Observations:

1. From this we can understand that 75% of the patients who has survived for more than 5 years are having axil_nodes around 5 and less than 10 in number.
2. Nearly 50% of patients who didnot survive are having axil_nodes in the range of [4 - 8]

```
a = df.loc[df['survival_status'] == 1]
print('Maximum number  of axil nodes present in a patient who survived
for more than 5 years after operation is:', max(a.axil_nodes))
print('Minimum number  of axil nodes present in a patient who survived
for more than 5 years after operation is:', min(a.axil_nodes))
print('*'*110)
b = df.loc[df['survival_status'] == 2]
print('Maximum number  of axil nodes present in a patient who survived
for less than 5 years after operation is:', max(b.axil_nodes))
print('Minimum number  of axil nodes present in a patient who survived
for less than 5 years after operation is:', min(b.axil_nodes))
```

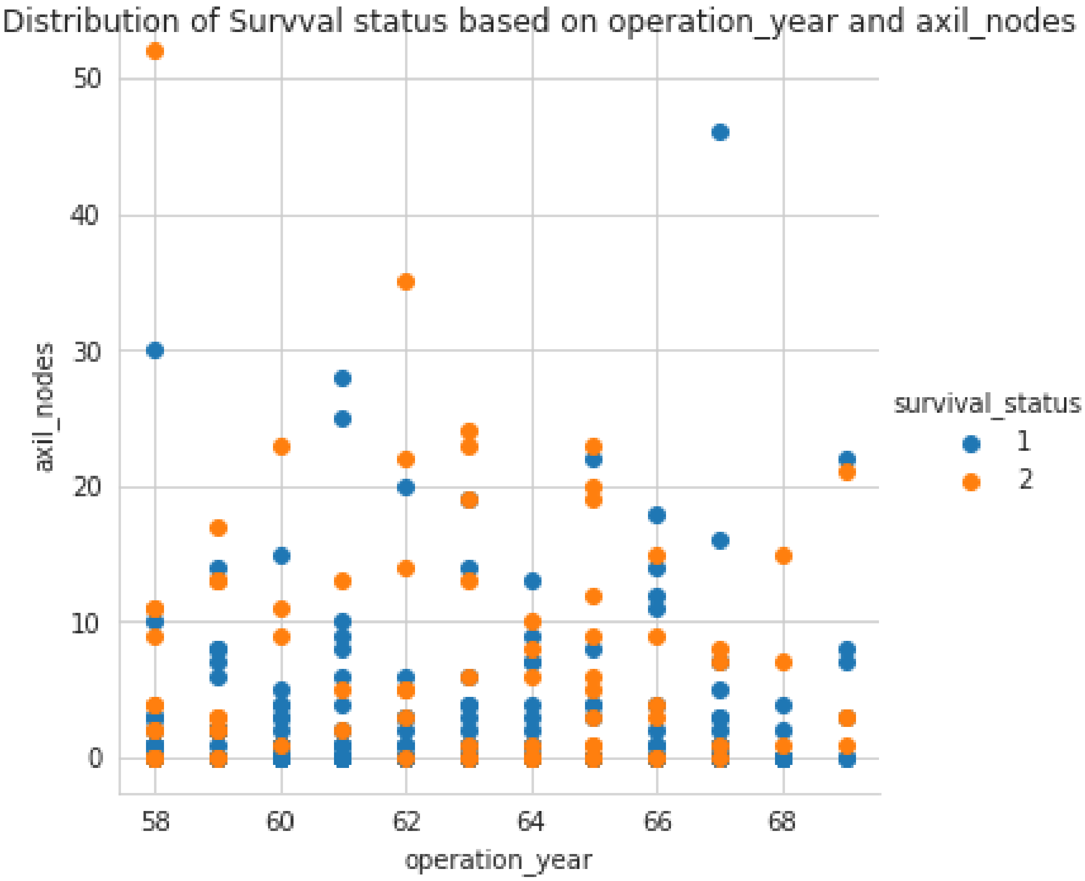
```
Maximum number  of axil nodes present in a patient who survived for
more than 5 years after operation is: 46
Minimum number  of axil nodes present in a patient who survived for
more than 5 years after operation is: 0
**********************************************************************
****************************************
Maximum number  of axil nodes present in a patient who survived for
less than 5 years after operation is: 52
Minimum number  of axil nodes present in a patient who survived for
less than 5 years after operation is: 0
```

Observation:

1. From this we can understand that survial status of patients is not only depending on the number of axil_nodes but there are some other reasons as well

Bi-variate Analysis

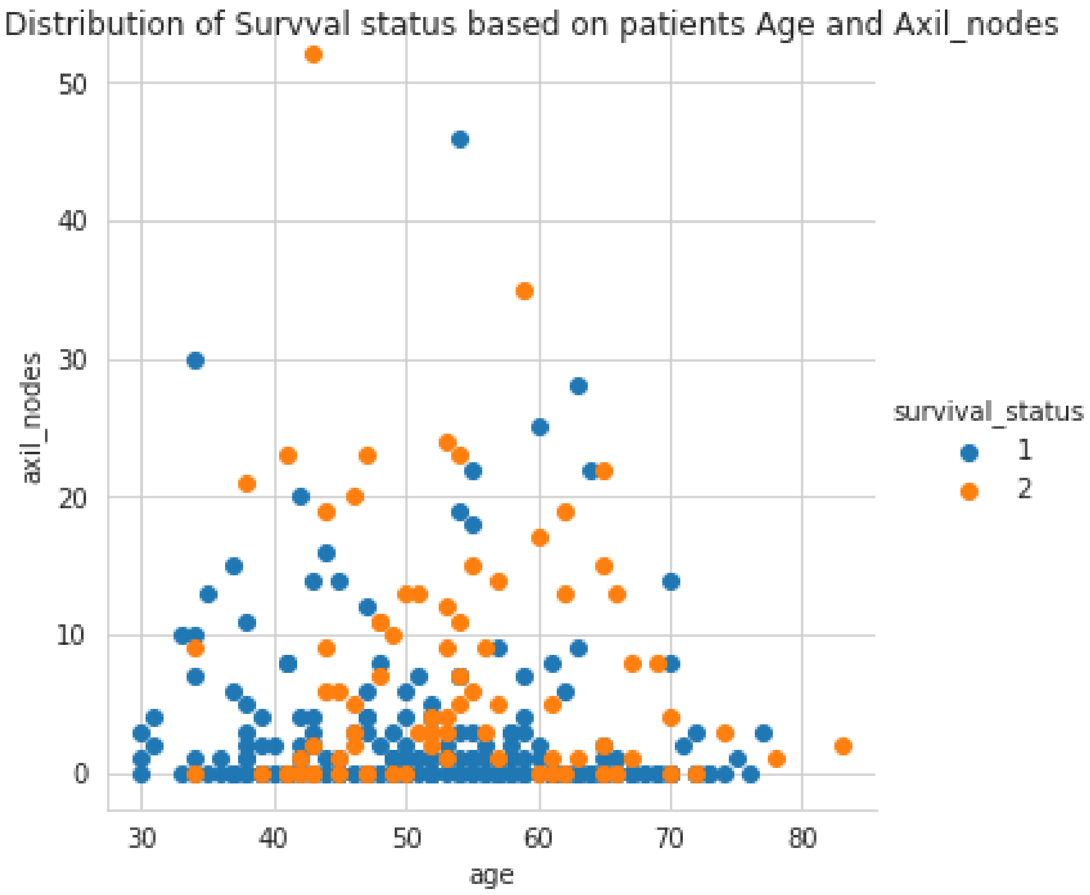
```
fg = sns.set_style("whitegrid")
fg = sns.FacetGrid(df,hue="survival_status",height = 5)\
     .map(plt.scatter,"operation_year","axil_nodes")\
     .add_legend()

fg.fig.suptitle('Distribution of Survval status based on 
operation_year and axil_nodes') # adding title
plt.show();
plt.show()
```



Observation:

1. From this above graph we cant decide survival status based on patients operation_year and number of axil_nodes as most of the values are overlapped

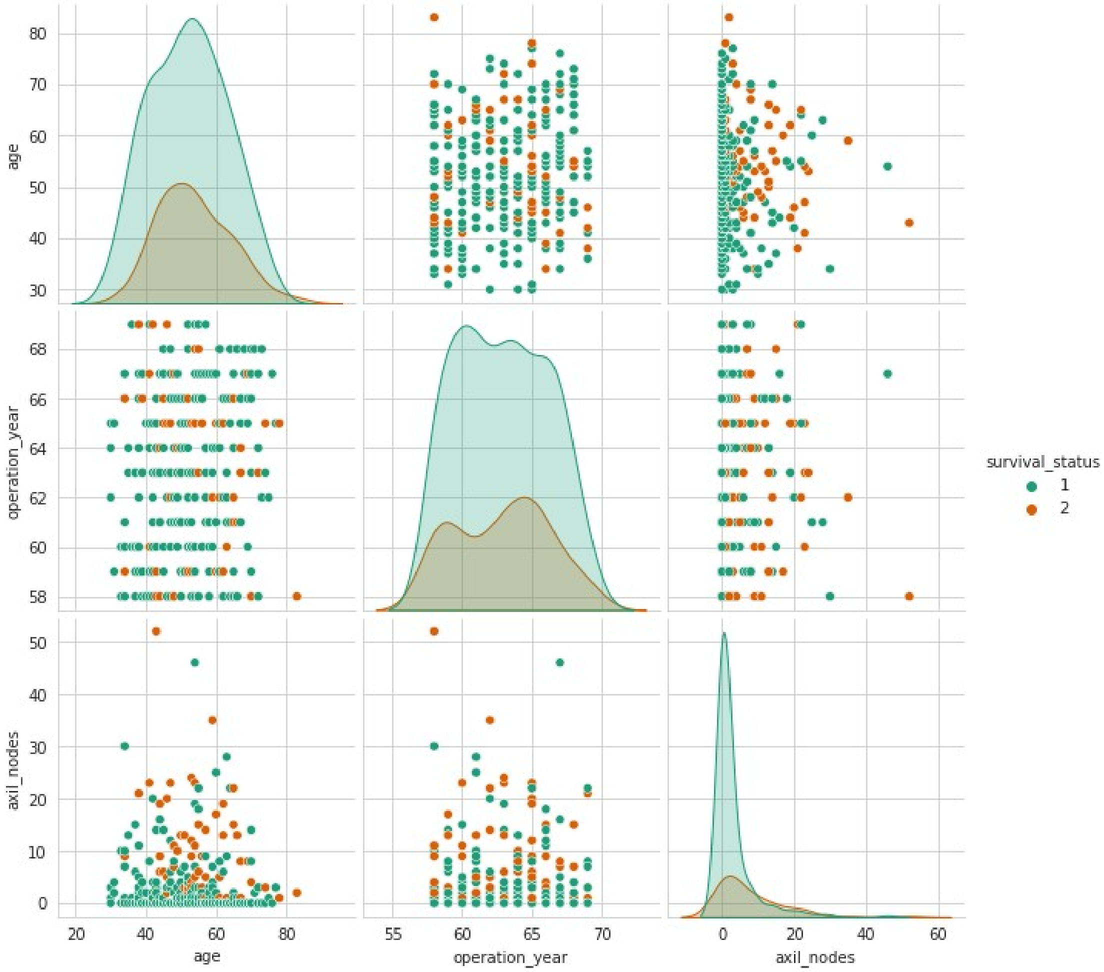
```
fg = sns.set_style("whitegrid")
fg = sns.FacetGrid(df,hue="survival_status",height = 5)\
     .map(plt.scatter,"age","axil_nodes")\
     .add_legend()

fg.fig.suptitle('Distribution of Survval status based on patients Age
and Axil_nodes') # adding title
plt.show();
```



Observations:

1. From this above graph we cant decide survival status based on patients age and number of axil_nodes as most of the values are overlapped
2. We can understand that based on the combination of any 2 features its not easy to decide survival rate of a patient

```
fg = sns.set_style("whitegrid")
fg = sns.pairplot(data = df, hue="survival_status", height = 3,
palette ='Dark2')

plt.show();
```

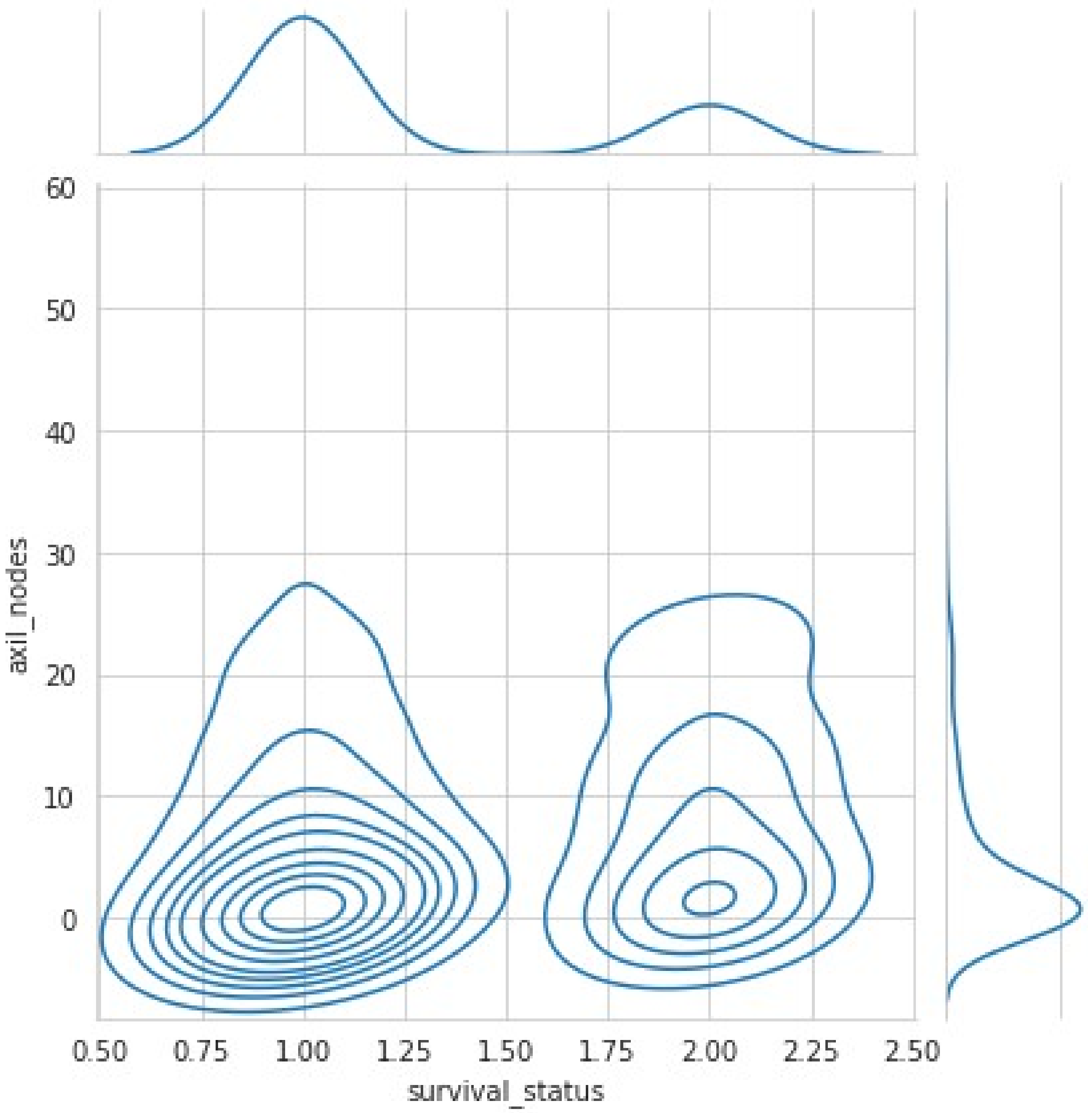


Observation:

1. From this we cant make any decision for the survival of the patient using the combination of any 2 features because all most all of them are overlapped with each other

Multivariate Probability density

```
sns.jointplot( data = df ,x = "survival_status", y = "axil_nodes",
kind="kde", palette = 'autumn')

plt.grid()
plt.show()
```



Observation:

1. First contour represents for **survival_status** = 1 and most of them are having axil_nodes around 0 in number
2. Second contour represents for **survival_status** = 2 and we can observe that only fewer number of patients are having axil_nodes = 0 in number and has died witin 5 years of operation

Overall Conclusions from EDA

From 2D plots

1. After the operations 74% of patients has survived even after 5 years of operation whereas 26% of patients has died within 5 years.

2. Number of people who has died within 5 years is significantly reduced in 1969 when compared to 1958.
3. Number of operations that are done on patients are very less in number in the years 1968 and 1969 when compared to all the other years.
4. Survival status has no relation with Age and Operation_year feature but patients who are having age in between 55 to 60 are having more survival rate when compared to people of other age groups

From Pair plots and Scatter plots

1. Age and Axil_nodes features can be considered as best features for the classification of survival rates of the patients

From CDF's and PDF's

1. Aroung 95% of patients are having age less than 70 years
2. Most of the patients are having number of axil_nodes less than 10% of the patients are having axil_nodes in the range of 40 to 50
3. Nearly 50% of the patients who died with 5 years of operation are having axil_nodes in thr range of 2 to 11

General conclusions

1. Average age of patients in our Data if 52
2. From all the above observations we can conclude that ***AXIL_NODES*** feature can be considered as one of the most important feature of all
3. Patients who has ***SURVIVAL_STATUS*** = 1 has lesser number of ***AXIL_NODES*** detected when compared to patients who have ***SURVIVAL_STATUS*** = 2 (who died within 5 years of their operation)
4. Except ***AXIL_NODES*** feature all the other features will not play that much vital role in determining whether a patient would survive or not
5. Number of people who has undergone operation is very much less in 1969 when compared to 1958, this may be due to awareness about the disease and it was cured without requiring any operation