Exploratory Data Analysis - Haberman Dataset

INDEX

OBJECTIVES

1.The primary objective is to find the survival of patients after the treatment of breast cancer surgically.

- Patients who survived more than or equal to 5 years.
- Patients who survived less than 5 years.

INTRODUCTION TO DATASET

- The Haberman dataset is a case study that was performed 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.

- There are following fields in the dataset:
  - Features
  - Age of patient at time of operation
  - Patient’s year of operation
  - Number of positive auxillary nodes detected
  - Survival status (class)
  - 1 - the patients who survived more than or equal to 5 years.
  - 2 - the patients who survived less than 5 years.

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

haberman = pd.read_csv(".\dataset\haberman.csv")
print(haberman)
```

```
# create DataFrame for the table

haberman_df = pd.DataFrame(haberman)
print(haberman_df.head())
print(haberman_df.tail())

shp = haberman_df.shape
col = haberman_df.columns

print(shp)
print(col)

     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]
   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
     age  year  nodes  status
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, 4)
Index(['age', 'year', 'nodes', 'status'], dtype='object')
```

Describe the dataset and show information related to dataset High level Statistics

```
description  = haberman_df.describe()
print(description)
haberman_df.info()

              age         year        nodes       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
<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   year    306 non-null    int64
 2   nodes   306 non-null    int64
 3   status  306 non-null    int64
dtypes: int64(4)
memory usage: 9.7 KB
```

Observation - : Total record = 306

- Minimum age at which Breat cancer encountered = 30
- Maximum age at which Breat cancer encountered = 83
- Mean age of Breast cancer patients = 52
- Maximum number of positive auxilliary nodes in which cancer cells were found was = 52.

Check for any erroneous value

```
redundant_data = haberman_df.isnull().sum()
print(redundant_data)
```

```
age       0
year      0
nodes     0
status    0
dtype: int64
```

Check for number of Datapoints for each status

```
haberman_df["status"].value_counts()
```

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

Observation - : Total record = 306

- Patients who survived more than or equal to five years is 225.
- Patients who survived less than 5 years is 81.
- Data imbalanced

Check for the presence of outliers

```
central_tendency_age = np.median(haberman_df["age"])
central_tendency_nodes = np.median(haberman_df["nodes"])
print(central_tendency_age)
print(central_tendency_nodes)
```

```
52.0
1.0
```

Observation - :

- check for the presence of any outliers in the column - 1. Ages 2. Nodes
- It seems that there is an outlier present in the Nodes column.
- Maximum nodes = 52 and central_tendency = 1.

Check Percentiles Quantiles and Inter Quantile Range.

```
quantiles = haberman_df.quantile([.1,.25,.5,.75,0.9,.95,.99],axis =0)
print(quantiles)
```

```
        age   year  nodes  status
0.10  38.00  58.00   0.00     1.0
0.25  44.00  60.00   0.00     1.0
0.50  52.00  63.00   1.00     1.0
0.75  60.75  65.75   4.00     2.0
0.90  67.00  67.00  13.00     2.0
0.95  70.00  68.00  19.75     2.0
0.99  75.95  69.00  29.90     2.0
```

Observation - :

- The percentiles show that the success rate of living above 5 years is more than or equal to 50 years of age.
- It shows that the probability to live a life after surgery is greater for patients of ages below 60 years.
- It also shows that the probabilty to live a life after surgery is greater for patients with lesser than 4 nodes

Univariate Analysis

Check for the Patient's age who are likely to encounter positive auxilliary nodes of Cancer cells

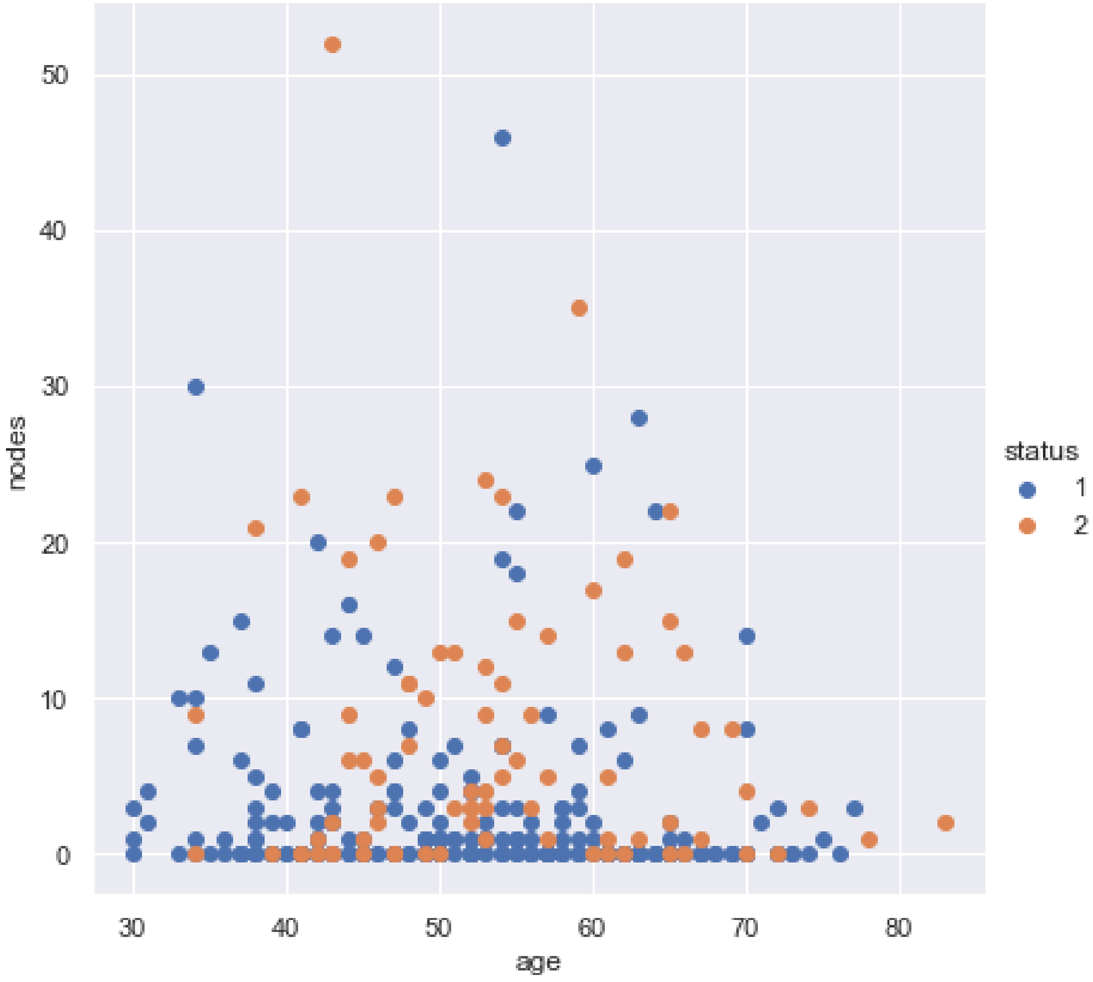
```
haberman['status'] = haberman['status'].map({1:'Success',2:'Failure'})
haberman.tail()
haberman.head()
print(haberman)
```

```
     age  year  nodes   status
0     30    64      1  Success
1     30    62      3  Success
2     30    65      0  Success
```

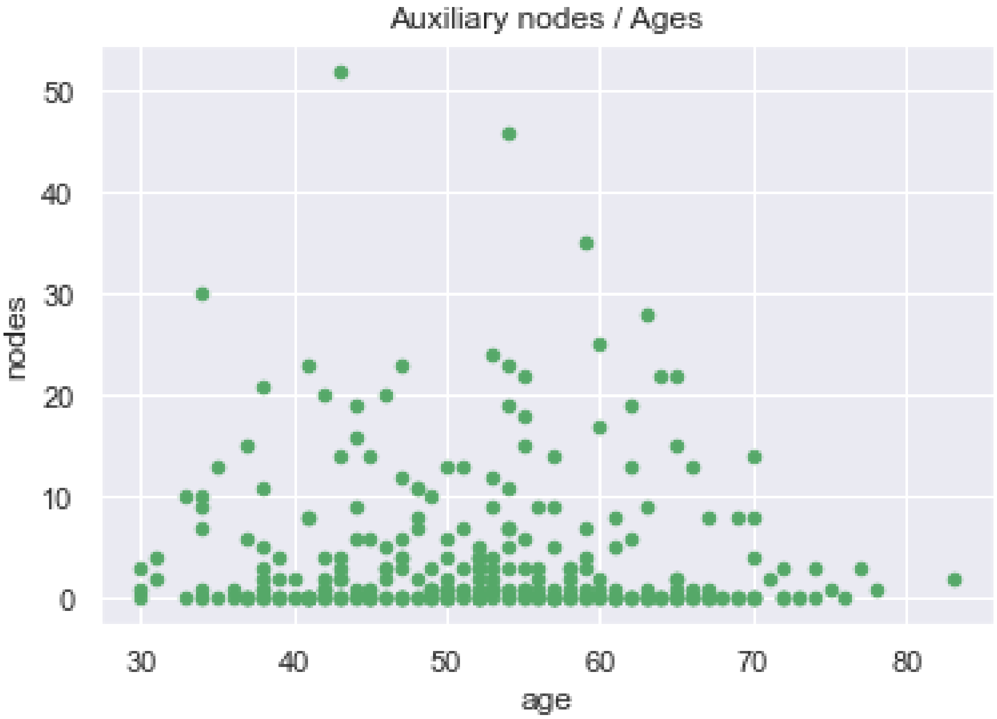
```
3     31    59      2  Success
4     31    65      4  Success
..   ...   ...    ...      ...
301   75    62      1  Success
302   76    67      0  Success
303   77    65      3  Success
304   78    65      1  Failure
305   83    58      2  Failure

[306 rows x 4 columns]
```

```
sns.set_theme(style = "darkgrid")
sns.FacetGrid(haberman, hue = "status",height =
6).map(plt.scatter,'age','nodes').add_legend()
plt.show()
```

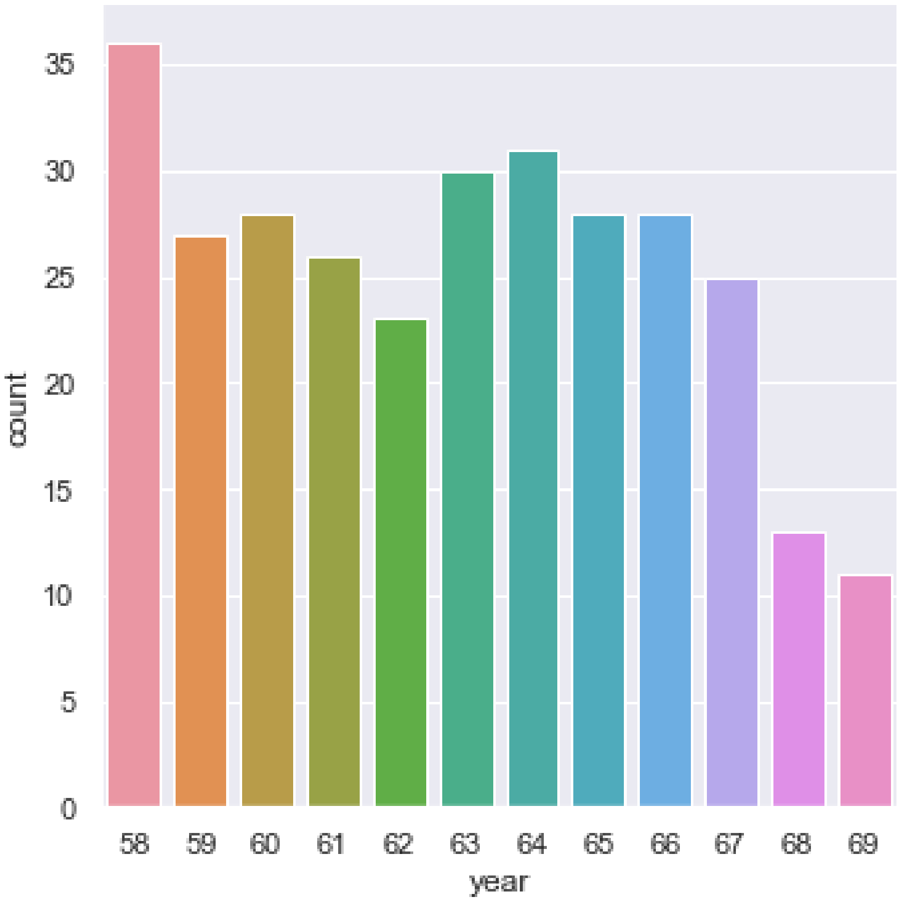


```
haberman_df.plot(kind = "scatter", x = "age",y = "nodes", color = "g")
plt.title("Auxiliary nodes / Ages")
plt.show()
```



```
sns.catplot(x="year", kind="count", data=haberman)
```

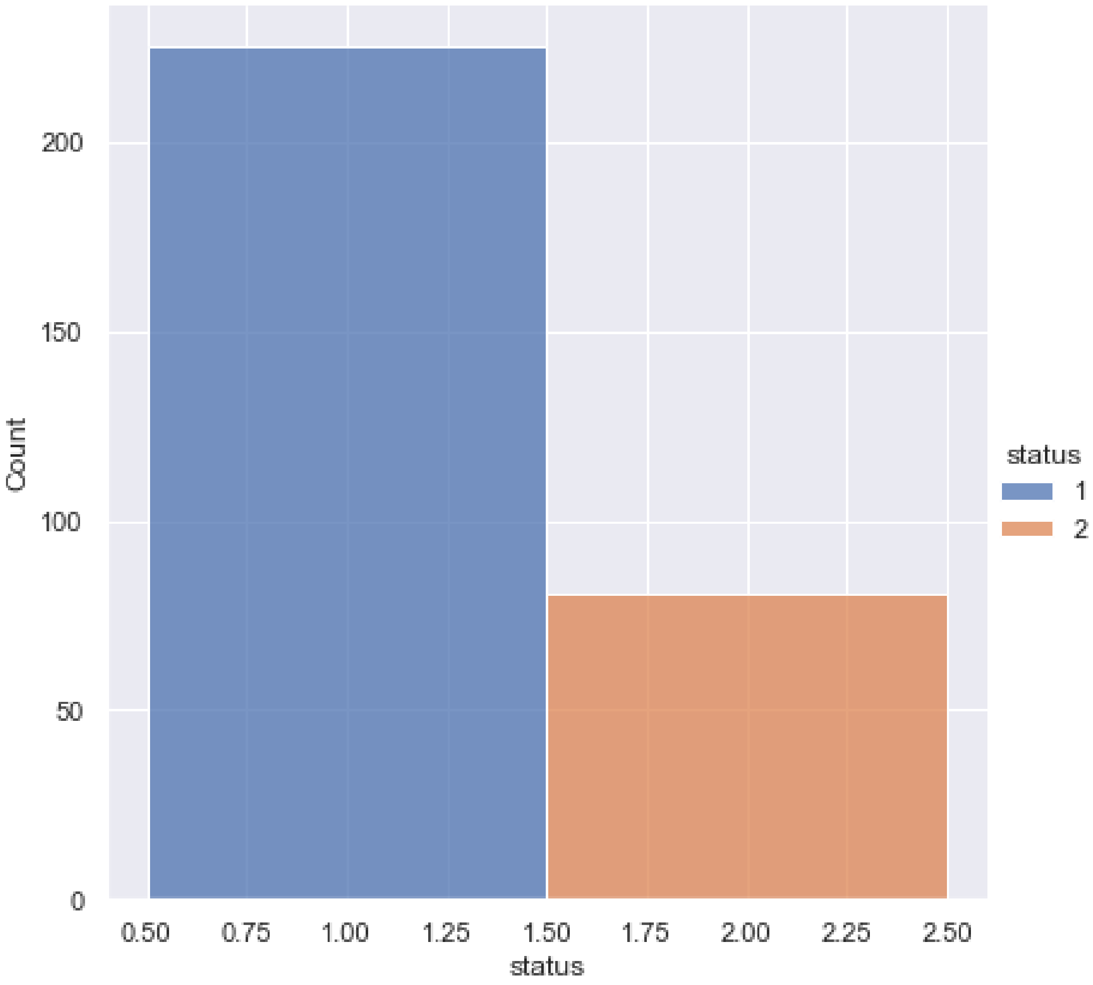
```
<seaborn.axisgrid.FacetGrid at 0x258ae4775b0>
```



OBSERVATION -:

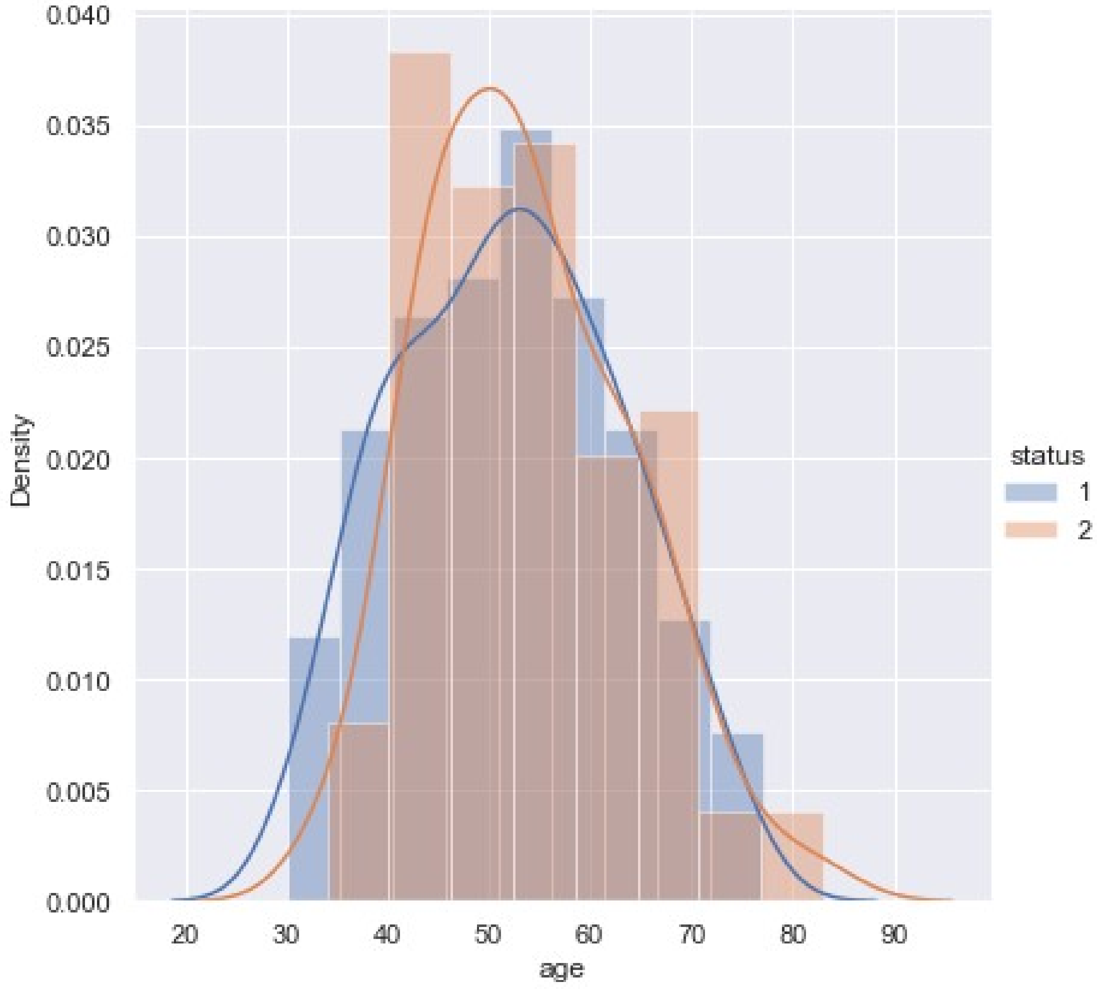
- This graph Clearly explains the maximum number of Case for Breast Cancer patients had been found in the year = 1958 and minimum nnumber of patients had been found in the year 1969.

```
sns.FacetGrid(haberman, hue = 'status' ,height =
6).map(sns.histplot,"status").add_legend()
plt.show()
```

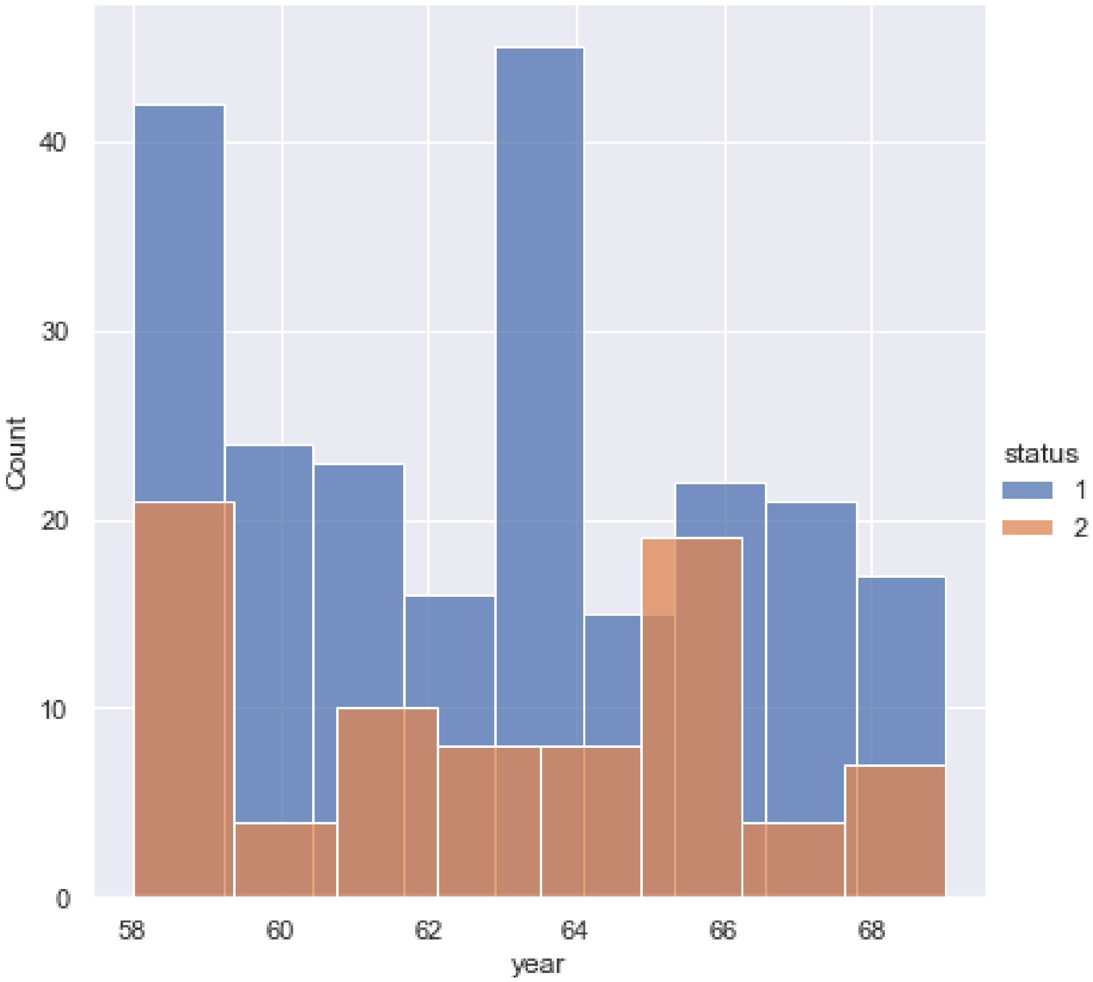


```
sns.FacetGrid(haberman, hue = 'status' ,height =
6).map(sns.distplot,"age").add_legend()
plt.show()
```

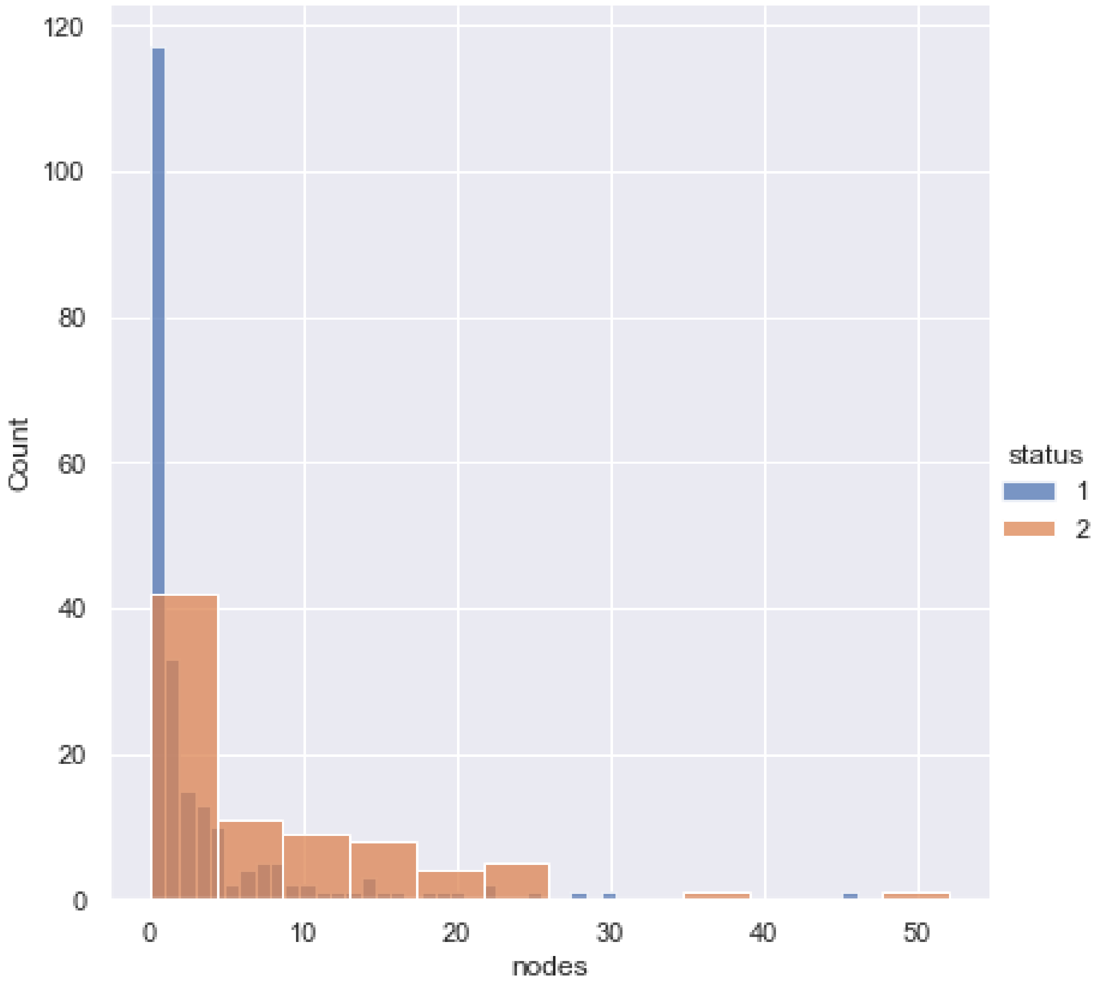
```
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\
distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your
code to use either `displot` (a figure-level function with similar
flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
C:\Users\DELL\anaconda3\lib\site-packages\seaborn\
distributions.py:2619: FutureWarning: `distplot` is a deprecated
function and will be removed in a future version. Please adapt your
code to use either `displot` (a figure-level function with similar
flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



```
sns.FacetGrid(haberman, hue = 'status' ,height =
6).map(sns.histplot,"year").add_legend()
plt.show()
```



```
sns.FacetGrid(haberman, hue = 'status' ,height =
6).map(sns.histplot,"nodes").add_legend()
plt.show()
```

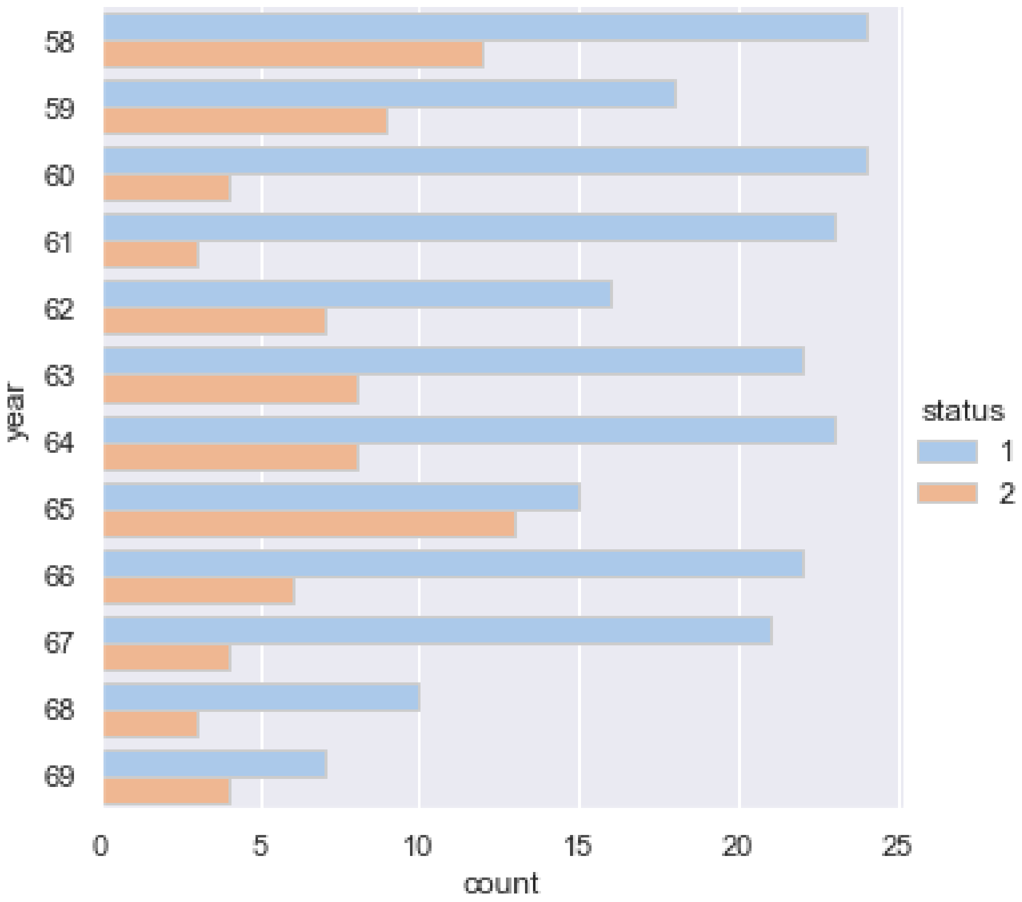


OBSERVATION -:

- From the above graphs, using ditribution plot and histogram plot we can't say much about the success and Failure, but in the graph with nodes it is clearly visible that success rate is inversely propotional to the number of nodes.

```
sns.catplot(y="year", hue="status", kind="count",
            palette="pastel", edgecolor=".8",
            data=haberman)
```

```
<seaborn.axisgrid.FacetGrid at 0x258b40b8520>
```

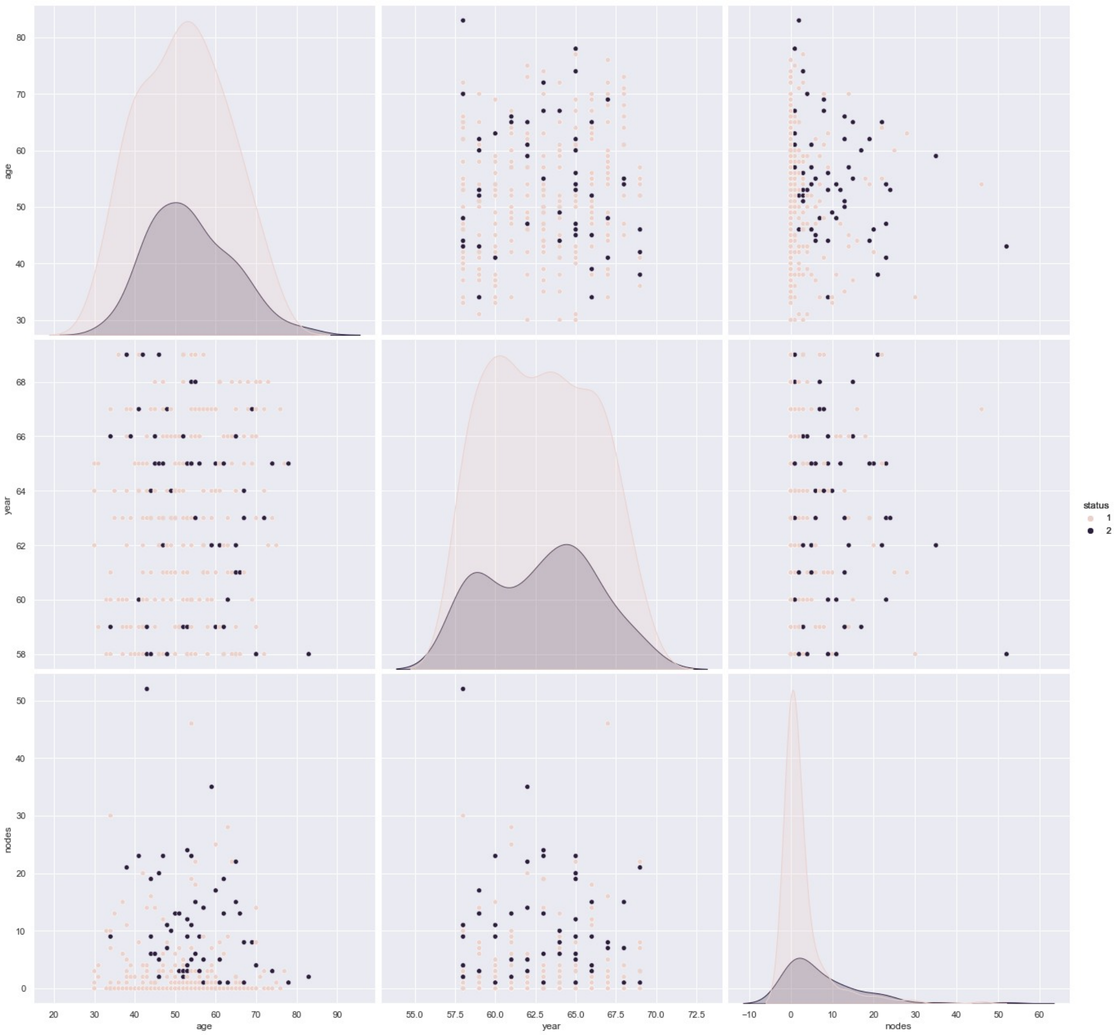


OBSERVATION -:

- Looking at the above graph it is pretty much clear that Year - 1961 was when Doctor's at the University of Chicago Billing's Hospital got much success in treating Breast Cancer Patients!

Multivariate

```
sns.pairplot(haberman,hue = "status",height = 6).add_legend()
plt.show()
```



Which year has highest number of cancer Surgery for BREAST CANCER

CUMULATIVE DISTRIBUTION FUNCTION

PROBABILITY DENSITY FUCNTION

```
# computing Pdf
counts, bin_edges = np.histogram(haberman["year"], bins = 10, density = True)
pdf = counts/(sum(counts))
print(counts)
print(pdf)

# At first generate the xlist pass it into the function and generate the ylist.
# And through these two lists plot the graph using the plot function.
# computing CDF

cdf = np.cumsum(pdf)
```

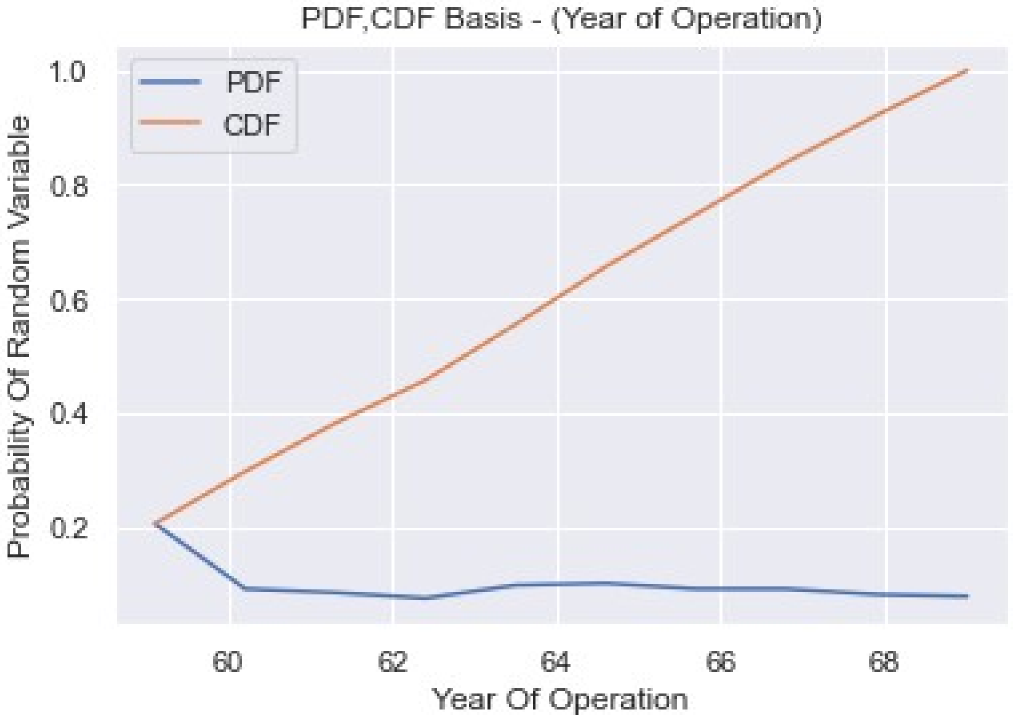
```
arr1, = plt.plot(bin_edges[1:],pdf)
arr2, = plt.plot(bin_edges[1:],cdf)
plt.legend([arr1,arr2],['PDF','CDF'])

plt.title("PDF,CDF Basis - (Year of Operation)")

plt.xlabel('Year Of Operation')
plt.ylabel('Probability Of Random Variable')

plt.show()
```

```
[0.18716578 0.08318479 0.07724302 0.06833036 0.08912656 0.09209745
 0.08318479 0.08318479 0.07427213 0.07130125]
[0.20588235 0.09150327 0.08496732 0.0751634  0.09803922 0.10130719
 0.09150327 0.09150327 0.08169935 0.07843137]
```



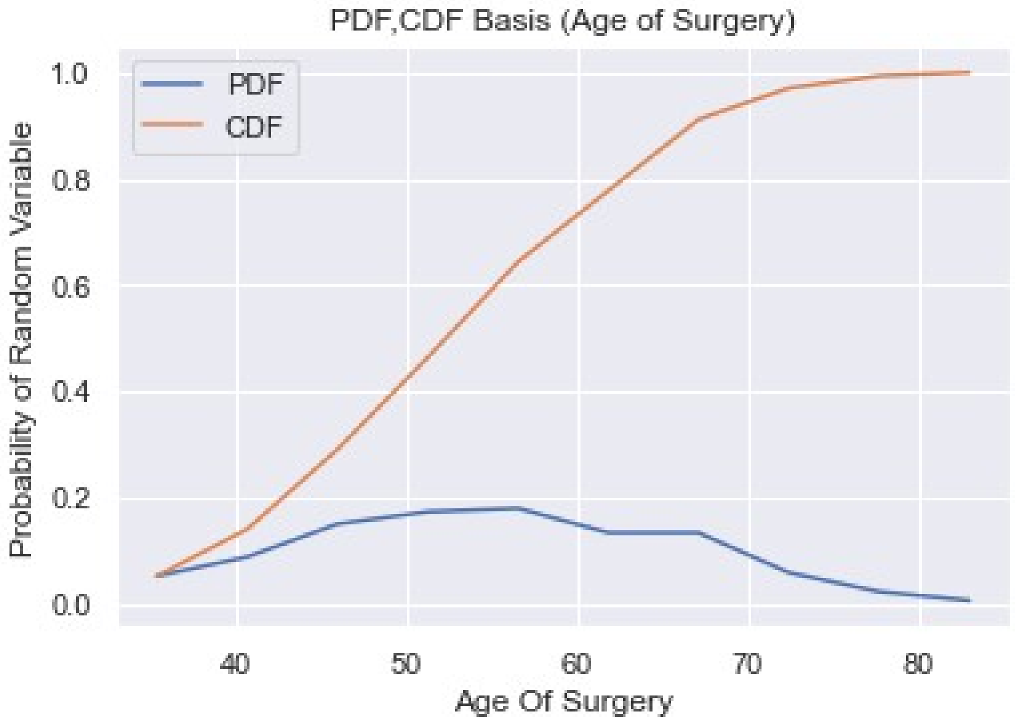
```
counts, bin_edges = np.histogram(haberman["age"], bins = 10, density =
True)
pdf = counts/(sum(counts))
print(counts)
print(pdf)

# computing CDF
cdf = np.cumsum(pdf)
arr1, = plt.plot(bin_edges[1:],pdf)
arr2, = plt.plot(bin_edges[1:],cdf)
plt.legend([arr1,arr2],['PDF','CDF'])
plt.title('PDF,CDF Basis (Age of Surgery)')

plt.xlabel('Age Of Surgery')
plt.ylabel('Probability of Random Variable')

plt.show()
```

```
[0.00986558 0.01664817 0.02836355 0.03267974 0.03391294 0.02528055
 0.02528055 0.01109878 0.00431619 0.0012332 ]
[0.05228758 0.08823529 0.1503268  0.17320261 0.17973856 0.13398693
 0.13398693 0.05882353 0.02287582 0.00653595]
```



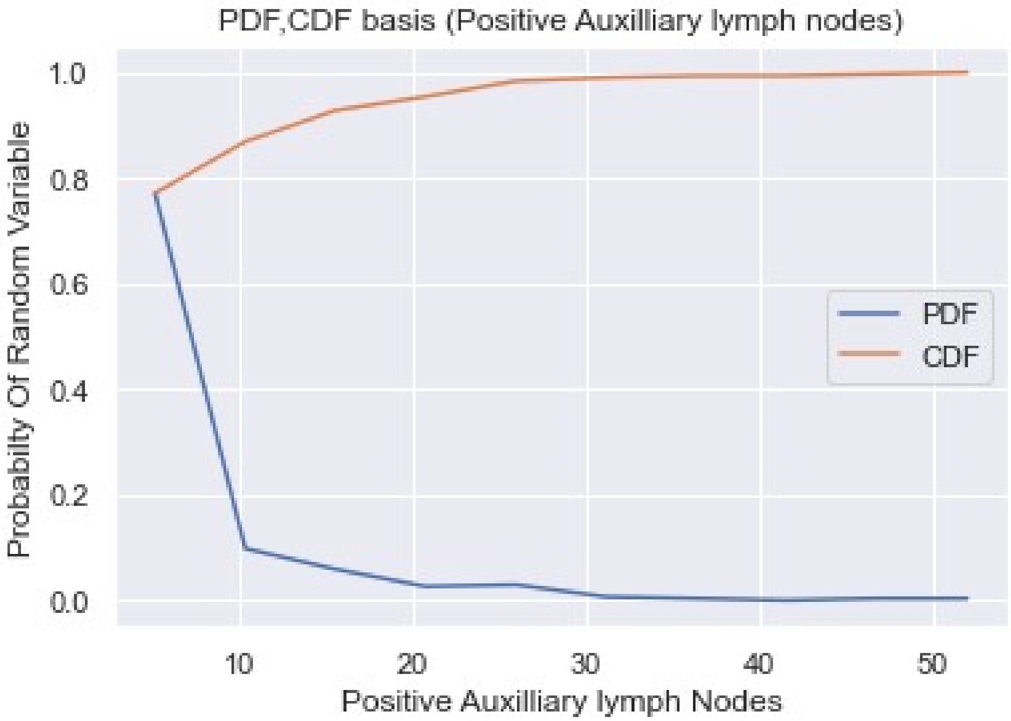
```
counts, bin_edges = np.histogram(haberman["nodes"], bins = 10, density
= True)
pdf = counts/(sum(counts))
print(counts)
print(pdf)


# computing CDF
cdf = np.cumsum(pdf)
arr1, = plt.plot(bin_edges[1:],pdf)
arr2, = plt.plot(bin_edges[1:],cdf)
plt.legend([arr1,arr2],['PDF','CDF'])

plt.title("PDF,CDF basis (Positive Auxilliary lymph nodes)")
plt.xlabel('Positive Auxilliary lymph Nodes')
plt.ylabel('Probabilty Of Random Variable')

plt.show()
```

```
[0.14831574 0.0188537  0.01131222 0.00502765 0.00565611 0.00125691
 0.00062846 0.         0.00062846 0.00062846]
[0.77124183 0.09803922 0.05882353 0.02614379 0.02941176 0.00653595
 0.00326797 0.         0.00326797 0.00326797]
```



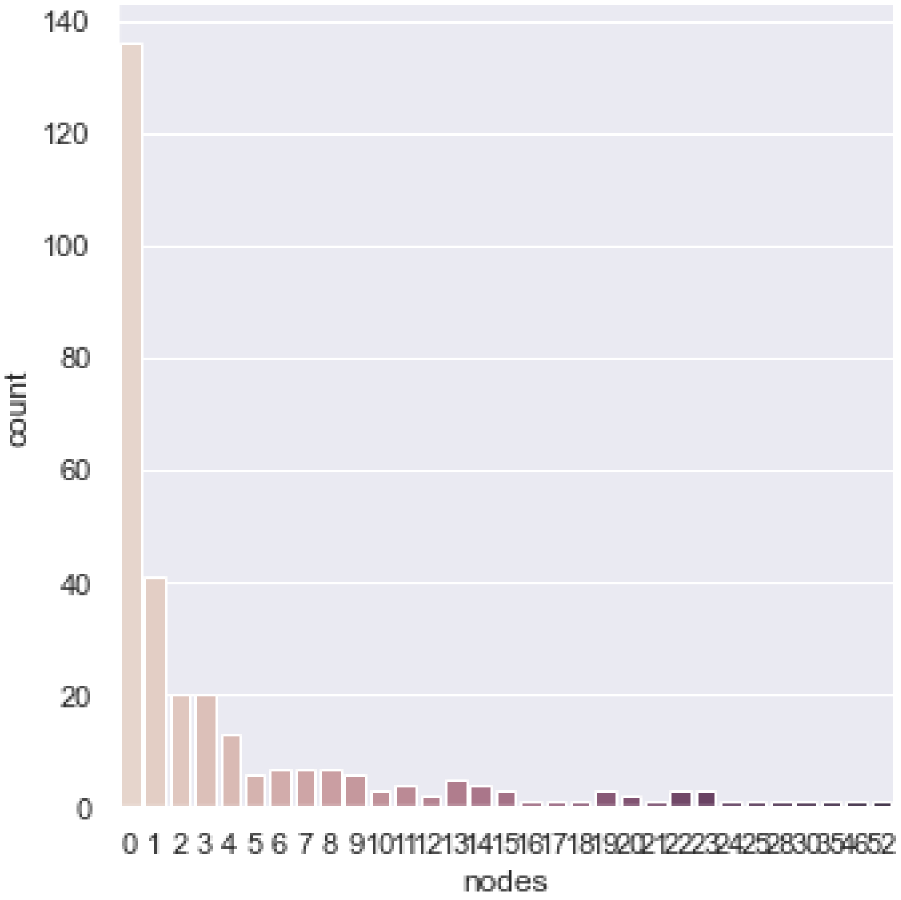
OBSERVATION -:

- Cumulative distribution fuction shows the percentage of patients cumulatively for Age, Nodes and Year it is between 0 to 100%.
- Pdf shows the Probability lieing between 0 to 1 for cumulative data, exmple - :The probability that a woman will encounter breast cancer is between the age of 50 to 60 years.
- The diffrentiation of CDF will give PDF and so Integration will give back the CDF.

check the data for individuals encountering presence of Cancer cells in auxilliary nodes

```
sns.catplot(x ="nodes", kind = "count", palette = "ch:.10", data = 
haberman)
```

```
<seaborn.axisgrid.FacetGrid at 0x1331a23a280>
```

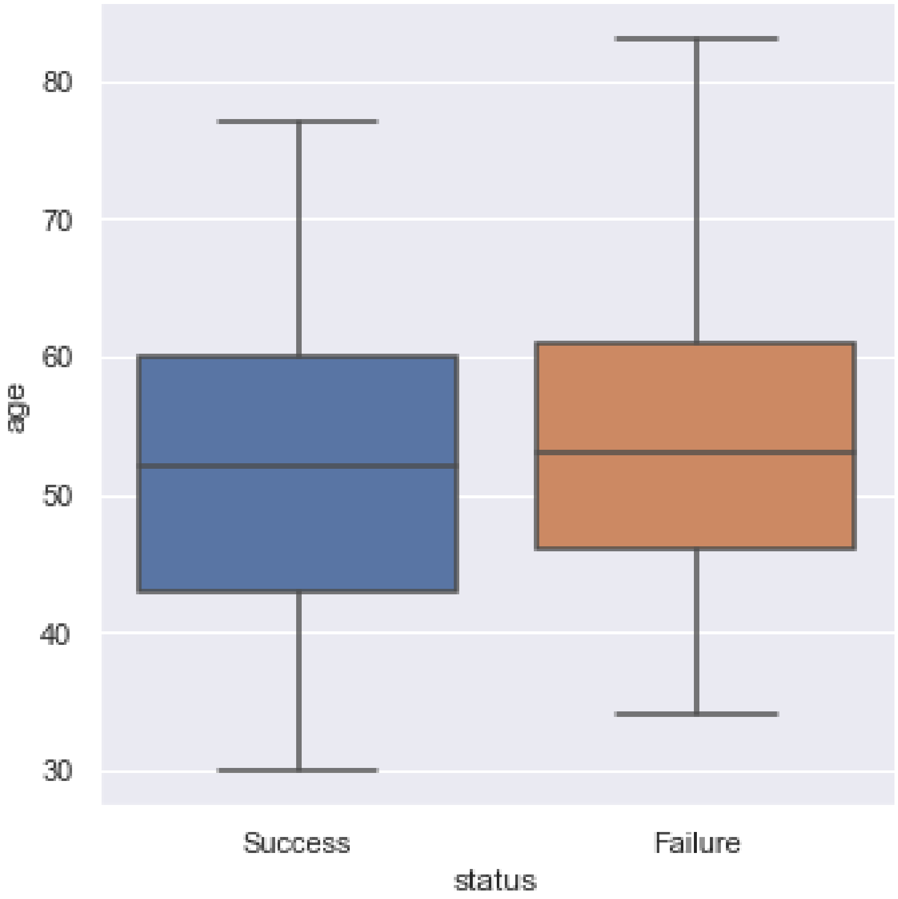


OBSERVATION -:

- This barplot in categorical plot gives the detail for the count of number of patients who have encountered Auxilliary nodes
- It is clearly visible that most of the patients have not been found with cancer spread in Auxiliary nodes which is approx 135

```
sns.catplot(data=haberman, orient="v", kind="box",x = 'status', y =
'age')
```

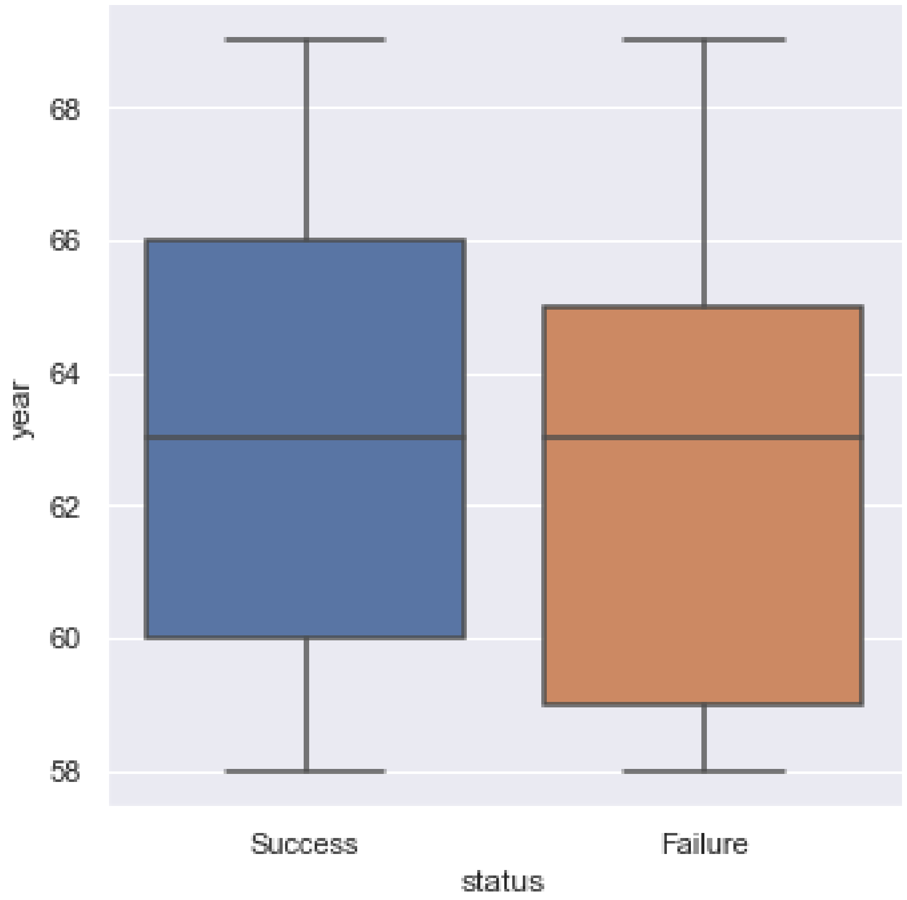
```
<seaborn.axisgrid.FacetGrid at 0x1331a63e9a0>
```



OBSERVATION -: The Success rate is maximum for the age group of women between 44 to 46 as clearly visible from the Boxplot

```
sns.catplot(data=haberman, orient="v", kind="box",x = 'status', y = 'year')
```

```
<seaborn.axisgrid.FacetGrid at 0x1331a4c1a00>
```

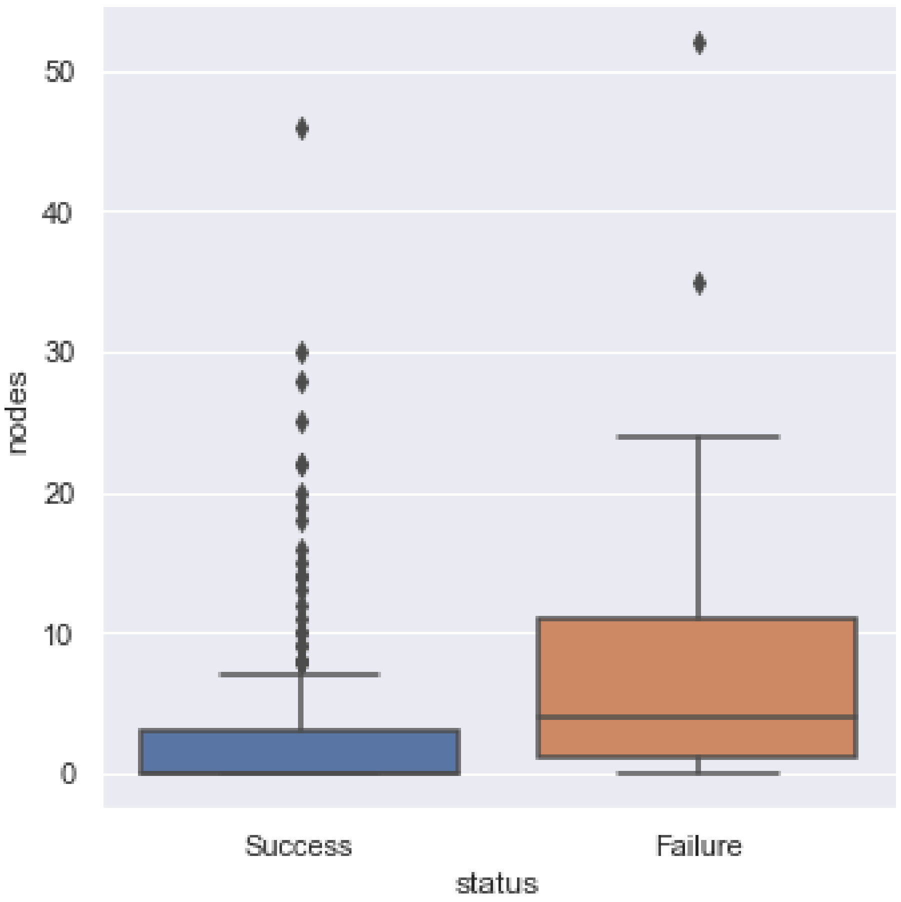


OBSERVATION -:

- The success rate was maximum in the year 1965 to 1966, as clearly visible from the boxplot.

```
sns.catplot(data=haberman, orient="v", kind="box",x = 'status', y = 
'nodes')
```

```
<seaborn.axisgrid.FacetGrid at 0x1331a50d430>
```



OBSERVATION -:

- Here it is clearly visible that the success rate is between 0 to 1 node greater than that there is less chance.

```
from statsmodels import robust
mad_age = robust.mad(haberman['age'])
mad_year = robust.mad(haberman['year'])
mad_nodes = robust.mad(haberman['nodes'])

print(mad_age)
print(mad_year)
print(mad_nodes)
```

```
11.860817748044816
4.447806655516806
1.482602218505602
```

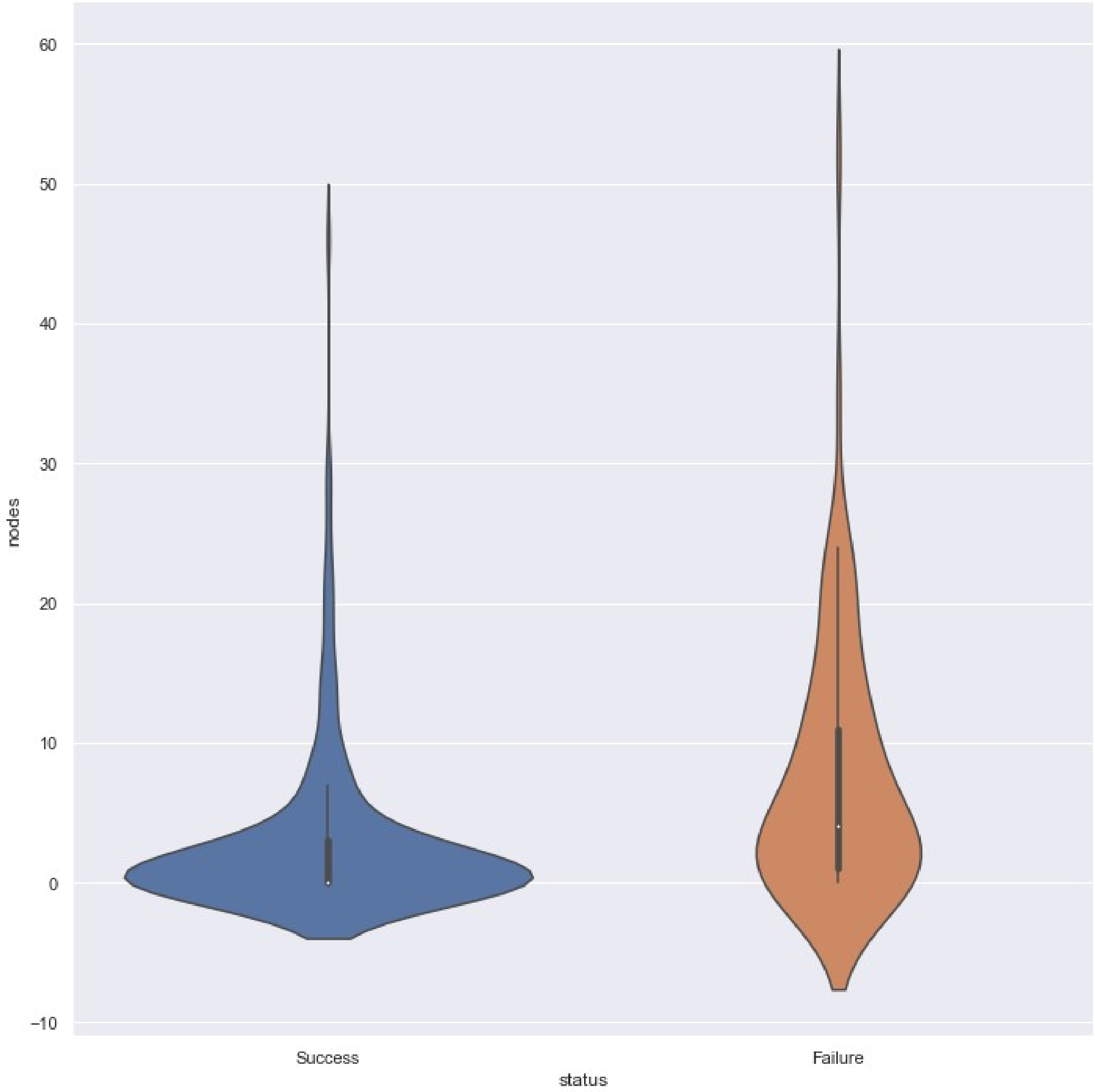
OBSERVATION -:

- The above results show the Median Standard Deviaton of the given data.

Violin Plot

```
sns.catplot(x='status', y='nodes', kind = "violin", data = haberman, height =10)
```

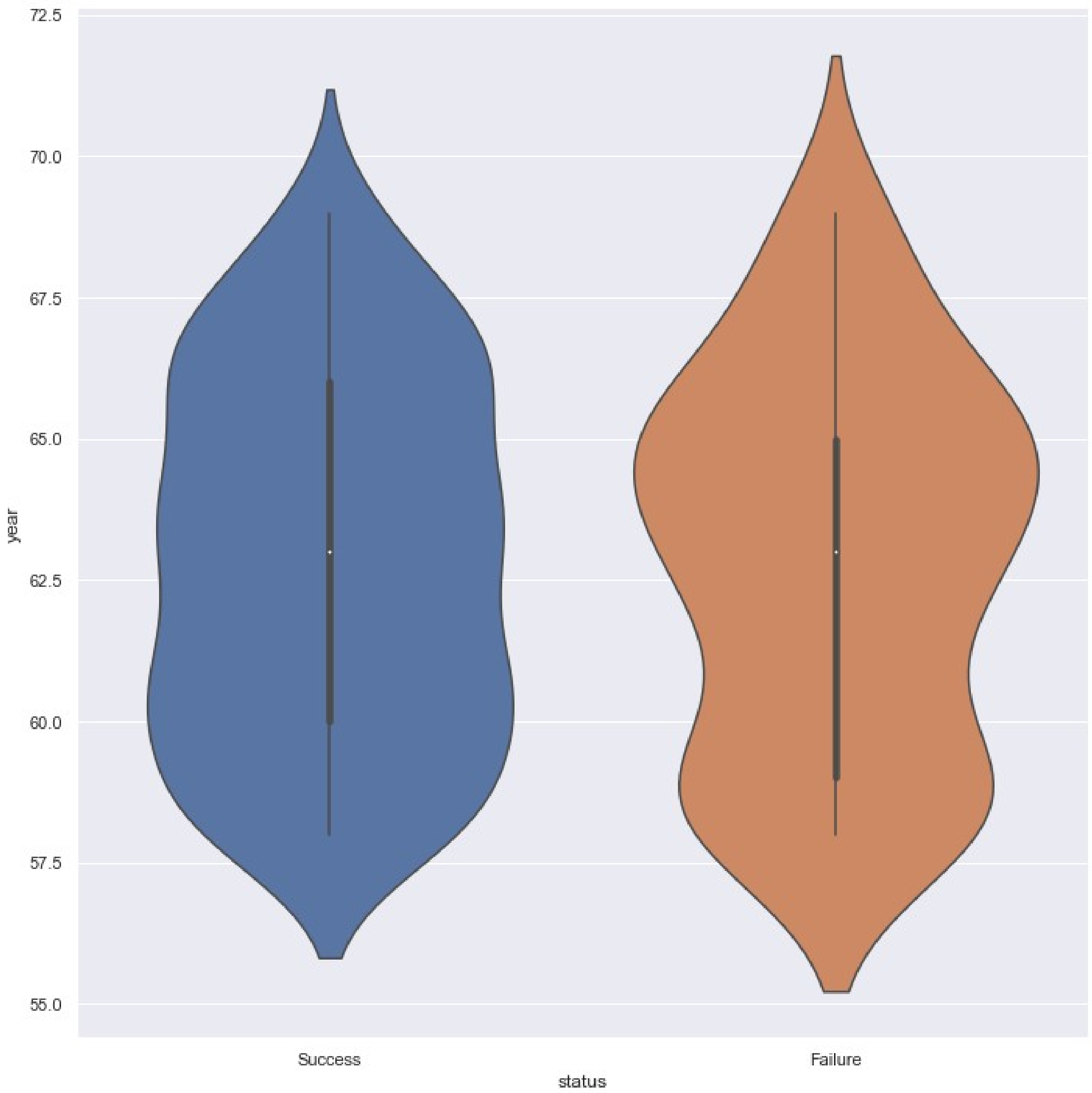
```
<seaborn.axisgrid.FacetGrid at 0x1331a9d3d30>
```



OBSERVATION -: The higher the number of nodes, the less chance of survival

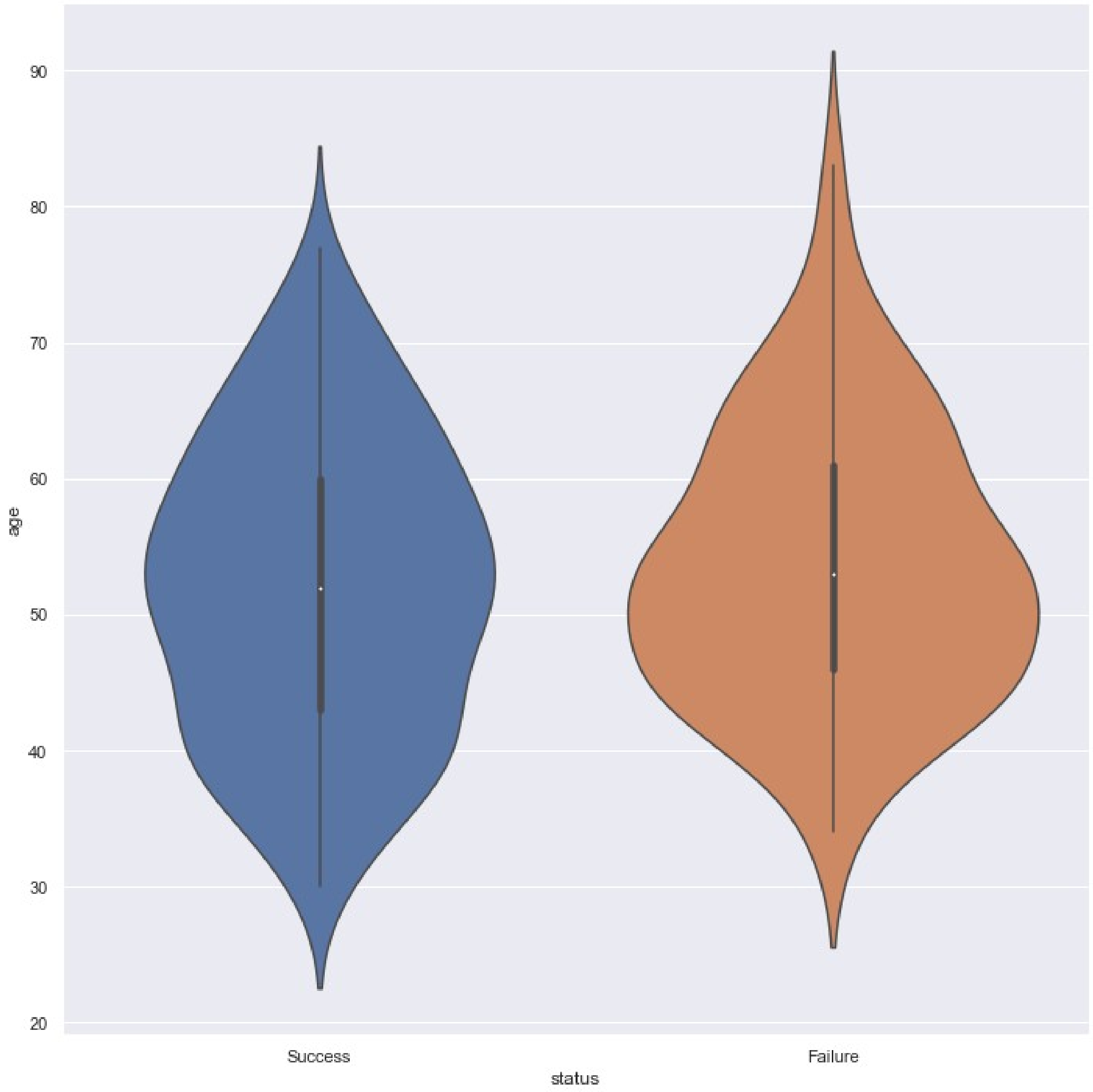
```
sns.catplot(x='status', y='year', kind = "violin", data = haberman, height = 10)
```

```
<seaborn.axisgrid.FacetGrid at 0x1331a6d88e0>
```



```
sns.catplot(x='status', y='age', kind = "violin", data = haberman,
height =10)
```

```
<seaborn.axisgrid.FacetGrid at 0x1331c23d6d0>
```



OBSERVATION -:

- The age band for a success in Breast cancer surgery is approximately less than 30.

CONCLUSION

1. From the above Analysis it can be clearly said that the survival chance of patients is inversely proportional to the number of nodes.
2. Also it is hard to tell about ages but we can conclude that lesser the age greater is the chance of survival.
3. Also it can be infered that in the year 1961 there was a huge success for the medical team at University of Chicago Billings Hospital.
4. The Success rate is maximum for the age group of women between 44 to 46.

OVERALL CONCLUSION OBSERVATION FROM HGIH LEVEL STATISTICS

Observation - Total record = 306

- Minimum age at which Breat cancer encountered = 30
- Maximum age at which Breat cancer encountered = 83
- Mean age of Breast cancer patients = 52
- Maximum number of positive auxilliary nodes in which cancer cells were found was = 52
- Patients who survived more than or equal to five years is 225.
- Patients who survived less than 5 years is 81.
- Data imbalanced

OBSERVATION FROM CENTRAL TENDENCY - :

1. check for the presence of any outliers in the column - 1. Ages 2. Nodes
2. It seems that there is an outlier present in the Nodes column.
3. Maximum nodes = 52 and central_tendency = 1.

OBSERVATION FROM PERCENTILES, QUANTILES

1. The percentiles show that the success rate of living above 5 years is more than or equal to 50 years of age.
2. It shows that the probability to live a life after surgery is greater for patients of ages below 60 years.
3. It also shows that the probabilty to live a life after surgery is greater for patients with lesser than 4 nodes

OBSERVATION FROM CATEGORICAL PLOT - (BAR GRAPH)

- The percentiles show that the success rate of living above 5 years is more than or equal to 50 years of age.
- It shows that the probability to live a life after surgery is greater for patients of ages below 60 years.
- It also shows that the probabilty to live a life after surgery is greater for patients with lesser than 4 nodes

OBSERVATION OF SUCCESS AND FAILURE FROM AUXILLIARY NODES

- From the above graphs, using ditribution plot and histogram plot we can't say much about the success and Failure,but in the graph with nodes it is clearly visible that success rate is inversely propotional to the number of nodes.

OBSERVATION OF SUCCESS AND FAILURE ON THE BASIS OF YEAR

- Looking at the above graph it is pretty much clear that Year - 1961 was when Doctor's at the University of Chicago Billing's Hospital got much success in treating Breast Cancer Patients!

OBSERVATION OF THE BASIS ON PDF/CDF

- Cumulative distribution fuction shows the probability of percentage of patients cumulatively for Age, Nodes and Year it is between 0 to 100%.
- Pdf shows the Probability lieing between 0 to 1 for cumulative data, exmple - :The probability that a woman will encounter breast cancer is between the age of 50 to 60 years.
- The diffrentiation of CDF will give PDF and so Integration will give back the CDF.

OBSERVATION OF NUMBER OF PATIENTS ON THE BASIS OF NODES.

- This barplot in categorical plot gives the detail for the count of number of patients who have encountered Auxilliary nodes.
- It is clearly visible that most of the patients have not been found with cancer spread in Auxiliary nodes which is approx 135.

OBSERVATION ON BOX and WHISKERS ON THE BASIS OF SUCCESS/FAILURE ACCORDING TO AGE.

- The Success rate is maximum for the age group of women between 44 to 46 as clearly visible from the Boxplot.
- The success rate was maximum in the year 1965 to 1966, as clearly visible from the boxplot.

OBSERVATION BASIS MEAN STANDARD DEVIATION

- The above results show the Median Standard Deviaton of the given data

OBSERVATION BASIS (VIOLIN PLOT)

- The age band for a success in Breast cancer surgery is approximately less than 30.