Exploratory Data Analysis of Haberman Dataset

Description : 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:
After referring to the given link(https://www.kaggle.com/gilsousa/habermans-survival-data-set):

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

Objective: To predict the patient's survival after 5 years based on the given features such as patient's age, year of operation and number of axiliary nodes.

In [2]:
```
# Importing the libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

In [ ]:
```
#loading the Haberman's data set into the pandas Dataframe
Haberman_df=pd.read_csv("haberman.csv")
```

Basic information:

In [14]:
```
#Number of data points and features
print(Haberman_df.shape)
Haberman_df.columns
```

```
(306, 4)
```

Out[14]:
```
Index(['Patient_age', 'op_year', 'Axil_nodes', 'Surv_status'], dtype='object')
```

Observation : The dataset has 306 rows and 4 columns

In [5]:
```
#Renaming the columns
Haberman_df.columns=['Patient_age','op_year','Axil_nodes','Surv_status']
Haberman_df
```

Out[5]:

	Patient_age	op_year	Axil_nodes	Surv_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
5	33	58	10	1
6	33	60	0	1
7	34	59	0	2
8	34	66	9	2
9	34	58	30	1
10	34	60	1	1
11	34	61	10	1
12	34	67	7	1
13	34	60	0	1
14	35	64	13	1
15	35	63	0	1
16	36	60	1	1
17	36	69	0	1
18	37	60	0	1
19	37	63	0	1
20	37	58	0	1
21	37	59	6	1
22	37	60	15	1
23	37	63	0	1
24	38	69	21	2
25	38	59	2	1
26	38	60	0	1
27	38	60	0	1
28	38	62	3	1
29	38	64	1	1
...	...	...	...	...
276	67	66	0	1
277	67	61	0	1
278	67	65	0	1
279	68	67	0	1
280	68	68	0	1
281	69	67	8	2
282	69	60	0	1
283	69	65	0	1
284	69	66	0	1
285	70	58	0	2
286	70	58	4	2
287	70	66	14	1
288	70	67	0	1
289	70	68	0	1
290	70	59	8	1
291	70	63	0	1
292	71	68	2	1
293	72	63	0	2
294	72	58	0	1
295	72	64	0	1
296	72	67	3	1
297	73	62	0	1
298	73	68	0	1
299	74	65	3	2
300	74	63	0	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 × 4 columns

In [6]:
```
print(Haberman_df["Surv_status"].value_counts())
```

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

Obervation :

- 225 patients have survived while 81 patients have died.
- The Surv_status column has significantly large number of survivors i.e. data is skewed in favour of 1.

In [3]:
```
Haberman_df.describe()
```

Out[3]:

	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

Observation:

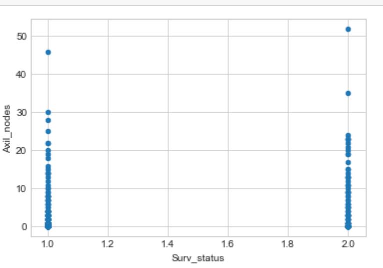
- Count : Total number of values present in respective columns.
- Mean: Sum of total values present divided by the count.
- Std: Amount of variation of a set of values.
- Min: The minimum value in the column.
- 25%: Gives the 25th percentile value.
- 50%: Gives the 50th percentile value.
- 75%: Gives the 75th percentile value.
- Max: The maximum value in the column

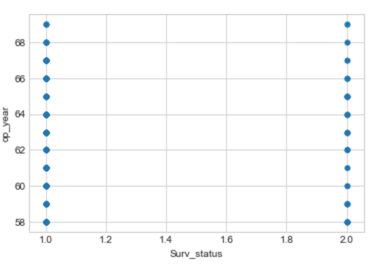
Bi-Variate Analysis

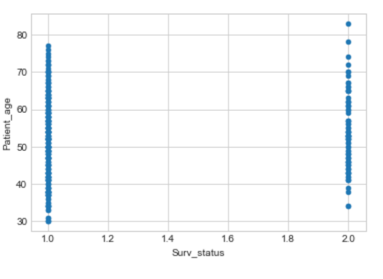
2-D Scatter plots

Scatter plots are data visualization method which is used for representing values of two different variables, one along x-axis and the other along y-axis.

In [7]:
```
# 2-D Scatter Plots
Haberman_df.plot(kind='scatter',y='Axil_nodes',x='Surv_status')
plt.show()
Haberman_df.plot(kind='scatter',y='op_year',x='Surv_status')
plt.show()
Haberman_df.plot(kind='scatter',y='Patient_age',x='Surv_status')
plt.show()
```





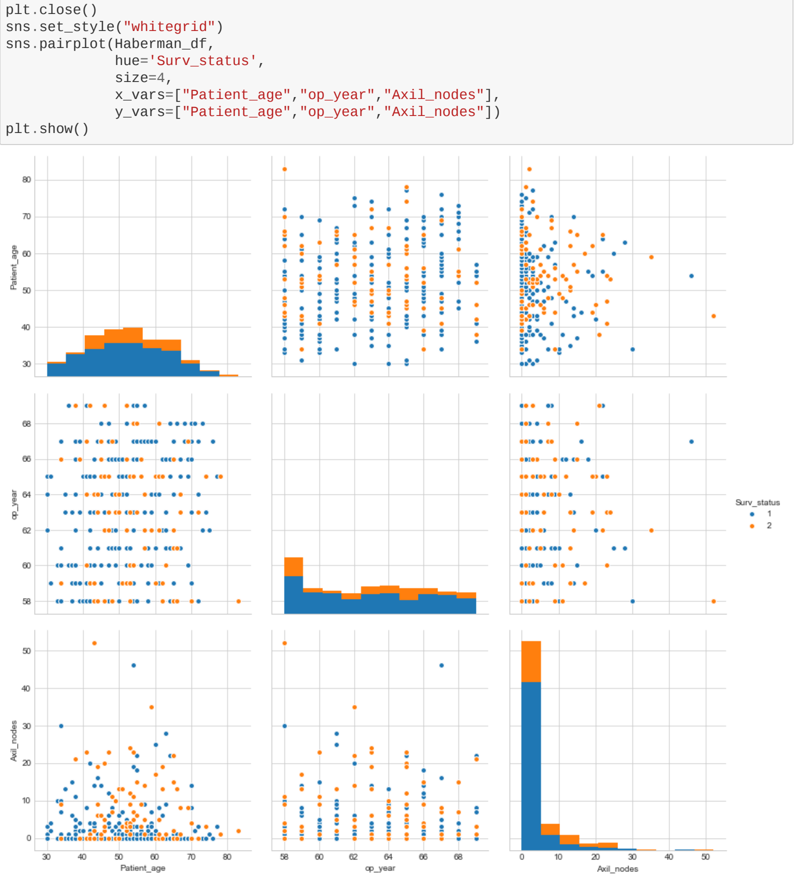


Observation: With 2-D scatter plots,no desicive conclusion can be drawn from the graph. Hence we move further with pair plots analysis.

Pair Plots

Pair plot allows us to see both distribution of single variables and relationships between two variables.

In [12]:
```
#Pair Plots
plt.close()
sns.set_style("whitegrid")
sns.pairplot(Haberman_df,
             hue='Surv_status',
             size=4,
             x_vars=["Patient_age","op_year","Axil_nodes"],
             y_vars=["Patient_age","op_year","Axil_nodes"])
plt.show()
```



Observation :

- Even though pair plots provide more information than 2-D scatter plots but the data points are overlapped together for all the pairs of features compared. Hence no conclusion can be drawn from this plot.

Univariate Analysis

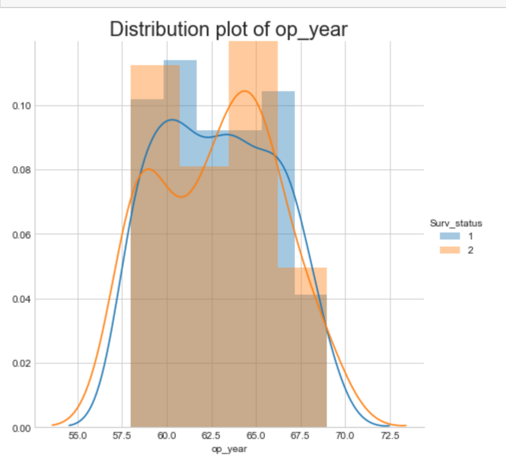
Univariate analysis means analysis of one variable or one feature and basically tells us how data in each feature is distributed.

In [9]:
```
# Univariate Analysis
sns.set_style('whitegrid')
sns.FacetGrid(Haberman_df,hue='Surv_status',size=6).map(sns.distplot,'Patient_age').add_lege
nd()
plt.title('Distribution of Patient_age',size=20)
plt.show()
```

Observation:

- A lot of overlapping can be observed in the plot but we can roughly observe that patients within the age of 40 are likely to survive and the survival rate decreases above the age of 45.
- Hence we cannot come to a conclusion just by considering the age of the patient.

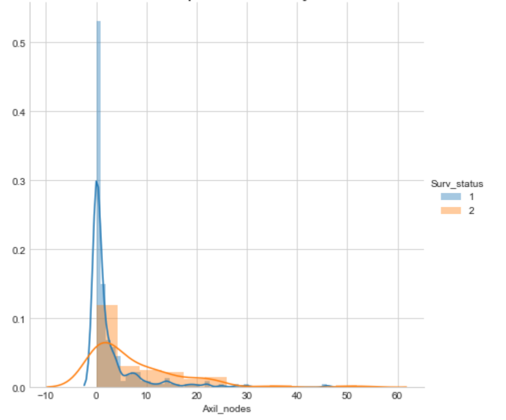
In [10]:
```
sns.set_style('whitegrid')
sns.FacetGrid(Haberman_df,hue='Surv_status',size=6).map(sns.distplot,'op_year').add_legend()
plt.title('Distribution plot of op_year',size=20)
plt.show()
```



Observation :

- Similar to patient's age parameter, significant overlapping of data with respect to the survival status of the patient can be seen in the op_year feature also.
- Hence no conclusion can be derived.

In [11]:
```
sns.set_style('whitegrid')
sns.FacetGrid(Haberman_df,hue='Surv_status',size=6).map(sns.distplot,'Axil_nodes').add_legen
d()
plt.title('Distribution plot of Axiliary nodes',size=20)
plt.show()
```



Observation :

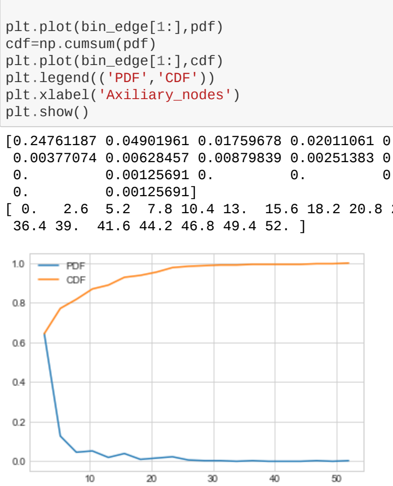
- Since the overlapping of data is less compared to other features, we can use the plot to get some insights into the patient's survival.
- Patients with less than two axiliary nodes are likely to survive and the survival rate decreases above 4 axiliary nodes. Very few patients who have 25 or above axiliary nodes survive.

Further analysis with Probability density fucntion(PDF) and Cumulative distribution function(CDF)

In [39]:
```
#Plotting the Probability Density Function(PDF) and Cumulative Distribution Function(CDF) of
axiliary node
counts,bin_edge=np.histogram(Haberman_df['Axil_nodes'],bins=20,density =True)
pdf=counts/(sum(counts))
print(counts)
print(bin_edge)

plt.plot(bin_edge[1:],pdf)
cdf=np.cumsum(pdf)
plt.plot(bin_edge[1:],cdf)
plt.legend(('PDF','CDF'))
plt.xlabel('Axiliary_nodes')
plt.show()
```

```
[0.24761187 0.04901961 0.01759678 0.02011061 0.00754148 0.01508296
 0.00377074 0.00628457 0.00879839 0.00251383 0.00125691 0.00125691
 0.         0.00125691 0.         0.         0.         0.00125691
 0.         0.00125691]
[ 0.   2.6  5.2  7.8 10.4 13.  15.6 18.2 20.8 23.4 26.  28.6 31.2 33.8
 36.4 39.  41.6 44.2 46.8 49.4 52. ]
```

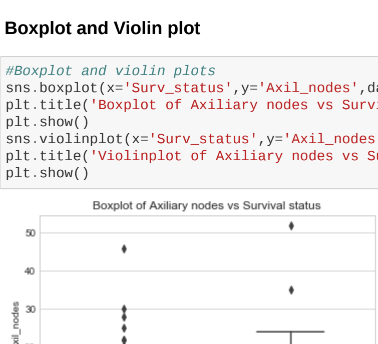


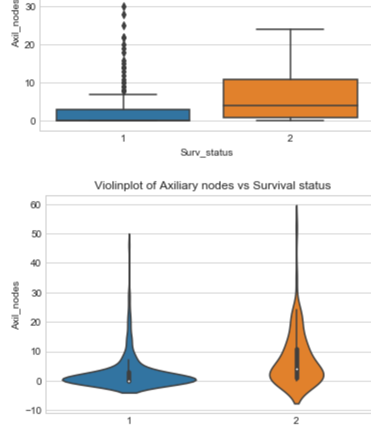
Observation :

- The CDF plot shows about 65% of the patients survived have axiliary nodes in the range of 0-4.

Boxplot and Violin plot

In [23]:
```
#Boxplot and violin plots
sns.boxplot(x='Surv_status',y='Axil_nodes',data=Haberman_df)
plt.title('Boxplot of Axiliary nodes vs Survival status')
plt.show()
sns.violinplot(x='Surv_status',y='Axil_nodes',data=Haberman_df)
plt.title('Violinplot of Axiliary nodes vs Survival status')
plt.show()
```





Observation:

- From the boxplot, it is clear that at the 75th percentile mark for the survived patients the axiliary nodes are less than 4 and the axiliary nodes is in the range of 3 to 11 for majority of the patients who have not survived.
- Similar observation can be concluded from the violin plot also.

Conclusion:

- Therefore out of all the features,axiliary nodes is the most insightful feature which helps in classifying the probability of patient's survival.
- Even though the data is imbalanced and there is some overlapping in the data, if the patient has less than 2 axiliary nodes then the patient is likely to survive and the survival isnt gauranteed with the absence of axiliary nodes as seen from the data.
- Survival rate is inversely proportional to number of axiliary nodes.