Exploratory Data Analysis of Haberman's DataSet

In [109]:
```
# Install the PyDrive wrapper & import libraries.
# This only needs to be done once in a notebook.
!pip install -U -q PyDrive
from pydrive.auth import GoogleAuth
from pydrive.drive import GoogleDrive
from google.colab import auth
from oauth2client.client import GoogleCredentials

# Authenticate and create the PyDrive client.
# This only needs to be done once in a notebook.
auth.authenticate_user()
gauth = GoogleAuth()
gauth.credentials = GoogleCredentials.get_application_default()
drive = GoogleDrive(gauth)

# Create & upload a text file.
uploaded = drive.CreateFile({'title': 'Sample file.txt'})
uploaded.SetContentString('Sample upload file content')
uploaded.Upload()
print('Uploaded file with ID {}'.format(uploaded.get('id')))
```

```
Uploaded file with ID 1zi619Ah7wg_BIKGm1tf1cEMsfvCxsiJJ
```

In [110]:
```
# Download a file based on its file ID.
file_id = '1YJzytNcpvROyRBGC_lRt_nUezp0dczae'
downloaded = drive.CreateFile({'id': file_id})
'Downloaded content "{}"'.format(downloaded.GetContentString())
# https://drive.google.com/open?id=1YJzytNcpvROyRBGC_lRt_nUezp0dczae
```

Out[110]:
```
'Downloaded content "age,operation_year,axil_nodes,status\n30,64,1,1
\n30,62,3,1\n30,65,0,1\n31,59,2,1\n31,65,4,1\n33,58,10,1\n33,60,0,1\n
34,59,0,2\n34,66,9,2\n34,58,30,1\n34,60,1,1\n34,61,10,1\n34,67,7,1\n3
4,60,0,1\n35,64,13,1\n35,63,0,1\n36,60,1,1\n36,69,0,1\n37,60,0,1\n37,
63,0,1\n37,58,0,1\n37,59,6,1\n37,60,15,1\n37,63,0,1\n38,69,21,2\n38,5
9,2,1\n38,60,0,1\n38,60,0,1\n38,62,3,1\n38,64,1,1\n38,66,0,1\n38,66,1
1,1\n38,60,1,1\n38,67,5,1\n39,66,0,2\n39,63,0,1\n39,67,0,1\n39,58,0,1
\n39,59,2,1\n39,63,4,1\n40,58,2,1\n40,58,0,1\n40,65,0,1\n41,60,23,2\n
41,64,0,2\n41,67,0,2\n41,58,0,1\n41,59,8,1\n41,59,0,1\n41,64,0,1\n41,
69,8,1\n41,65,0,1\n41,65,0,1\n42,69,1,2\n42,59,0,2\n42,58,0,1\n42,60,
1,1\n42,59,2,1\n42,61,4,1\n42,62,20,1\n42,65,0,1\n42,63,1,1\n43,58,5
2,2\n43,59,2,2\n43,64,0,2\n43,64,0,2\n43,63,14,1\n43,64,2,1\n43,64,3,
1\n43,60,0,1\n43,63,2,1\n43,65,0,1\n43,66,4,1\n44,64,6,2\n44,58,9,2\n
44,63,19,2\n44,61,0,1\n44,63,1,1\n44,61,0,1\n44,67,16,1\n45,65,6,2\n4
5,66,0,2\n45,67,1,2\n45,60,0,1\n45,67,0,1\n45,59,14,1\n45,64,0,1\n45,
68,0,1\n45,67,1,1\n46,58,2,2\n46,69,3,2\n46,62,5,2\n46,65,20,2\n46,6
2,0,1\n46,58,3,1\n46,63,0,1\n47,63,23,2\n47,62,0,2\n47,65,0,2\n47,61,
0,1\n47,63,6,1\n47,66,0,1\n47,67,0,1\n47,58,3,1\n47,60,4,1\n47,68,4,1
\n47,66,12,1\n48,58,11,2\n48,58,11,2\n48,67,7,2\n48,61,8,1\n48,62,2,1
\n48,64,0,1\n48,66,0,1\n49,63,0,2\n49,64,10,2\n49,61,1,1\n49,62,0,1\n
49,66,0,1\n49,60,1,1\n49,62,1,1\n49,63,3,1\n49,61,0,1\n49,67,1,1\n50,
63,13,2\n50,64,0,2\n50,59,0,1\n50,61,6,1\n50,61,0,1\n50,63,1,1\n50,5
8,1,1\n50,59,2,1\n50,61,0,1\n50,64,0,1\n50,65,4,1\n50,66,1,1\n51,59,1
3,2\n51,59,3,2\n51,64,7,1\n51,59,1,1\n51,65,0,1\n51,66,1,1\n52,69,3,2
\n52,59,2,2\n52,62,3,2\n52,66,4,2\n52,61,0,1\n52,63,4,1\n52,69,0,1\n5
2,60,4,1\n52,60,5,1\n52,62,0,1\n52,62,1,1\n52,64,0,1\n52,65,0,1\n52,6
8,0,1\n53,58,4,2\n53,65,1,2\n53,59,3,2\n53,60,9,2\n53,63,24,2\n53,65,
12,2\n53,58,1,1\n53,60,1,1\n53,60,2,1\n53,61,1,1\n53,63,0,1\n54,60,1
1,2\n54,65,23,2\n54,65,5,2\n54,68,7,2\n54,59,7,1\n54,60,3,1\n54,66,0,
1\n54,67,46,1\n54,62,0,1\n54,69,7,1\n54,63,19,1\n54,58,1,1\n54,62,0,1
\n55,63,6,2\n55,68,15,2\n55,58,1,1\n55,58,0,1\n55,58,1,1\n55,66,18,1
\n55,66,0,1\n55,69,3,1\n55,69,22,1\n55,67,1,1\n56,65,9,2\n56,66,3,2\n
56,60,0,1\n56,66,2,1\n56,66,1,1\n56,67,0,1\n56,60,0,1\n57,61,5,2\n57,
62,14,2\n57,64,1,2\n57,64,9,1\n57,69,0,1\n57,61,0,1\n57,62,0,1\n57,6
3,0,1\n57,64,0,1\n57,64,0,1\n57,67,0,1\n58,59,0,1\n58,60,3,1\n58,61,
1,1\n58,67,0,1\n58,58,0,1\n58,58,3,1\n58,61,2,1\n59,62,35,2\n59,60,0,
1\n59,63,0,1\n59,64,1,1\n59,64,4,1\n59,64,0,1\n59,64,7,1\n59,67,3,1\n
60,59,17,2\n60,65,0,2\n60,61,1,1\n60,67,2,1\n60,61,25,1\n60,64,0,1\n6
1,62,5,2\n61,65,0,2\n61,68,1,2\n61,59,0,1\n61,59,0,1\n61,64,0,1\n61,6
5,8,1\n61,68,0,1\n61,59,0,1\n62,59,13,2\n62,58,0,2\n62,65,19,2\n62,6
2,6,1\n62,66,0,1\n62,66,0,1\n62,58,0,1\n63,60,1,2\n63,61,0,1\n63,62,
0,1\n63,63,0,1\n63,63,0,1\n63,66,0,1\n63,61,9,1\n63,61,28,1\n64,58,0,
1\n64,65,22,1\n64,66,0,1\n64,61,0,1\n64,68,0,1\n65,58,0,2\n65,61,2,2
\n65,62,22,2\n65,66,15,2\n65,58,0,1\n65,64,0,1\n65,67,0,1\n65,59,2,1
\n65,64,0,1\n65,67,1,1\n66,58,0,2\n66,61,13,2\n66,58,0,1\n66,58,1,1\n
66,68,0,1\n67,64,8,2\n67,63,1,2\n67,66,0,1\n67,66,0,1\n67,61,0,1\n67,
65,0,1\n68,67,0,1\n68,68,0,1\n69,67,8,2\n69,60,0,1\n69,65,0,1\n69,66,
0,1\n70,58,0,2\n70,58,4,2\n70,66,14,1\n70,67,0,1\n70,68,0,1\n70,59,8,
1\n70,63,0,1\n71,68,2,1\n72,63,0,2\n72,58,0,1\n72,64,0,1\n72,67,3,1\n
73,62,0,1\n73,68,0,1\n74,65,3,2\n74,63,0,1\n75,62,1,1\n76,67,0,1\n77,
65,3,1\n78,65,1,2\n83,58,2,2\n"'
```

In [0]:
```
# Ignoring the warnings
import warnings
warnings.filterwarnings('ignore')
```

In [0]:
```
# Getting the content from the csv file
downloaded.GetContentFile('haberman.csv')
```

In [113]:
```
# Checking whether csv file is loaded or not
!ls
```

```
adc.json  haberman.csv  sample_data
```

In [0]:
```
# Importing the required libraries needed for the analysis
import numpy as py
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb
sb.set()
```

In [115]:
```
# Using pandas creating a data frame
df = pd.read_csv('haberman.csv')
df.head()
```

Out[115]:

	age	operation_year	axil_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

In [116]:
```
# Colums present in the data set
df.columns
```

Out[116]:
```
Index(['age', 'operation_year', 'axil_nodes', 'status'], dtype='objec
t')
```

In [117]:
```
# Finding the number of datapoins,feaatures
df.shape
```

Out[117]:
```
(306, 4)
```

Observation:

- **306 is the row count which indiactes the number of patients and 4 is the column count which indicates the variable count**

In [118]:
```
# Count of similar kind of data in the patient status
df['status'].value_counts()
```

Out[118]:
```
1    225
2     81
Name: status, dtype: int64
```

Observation:

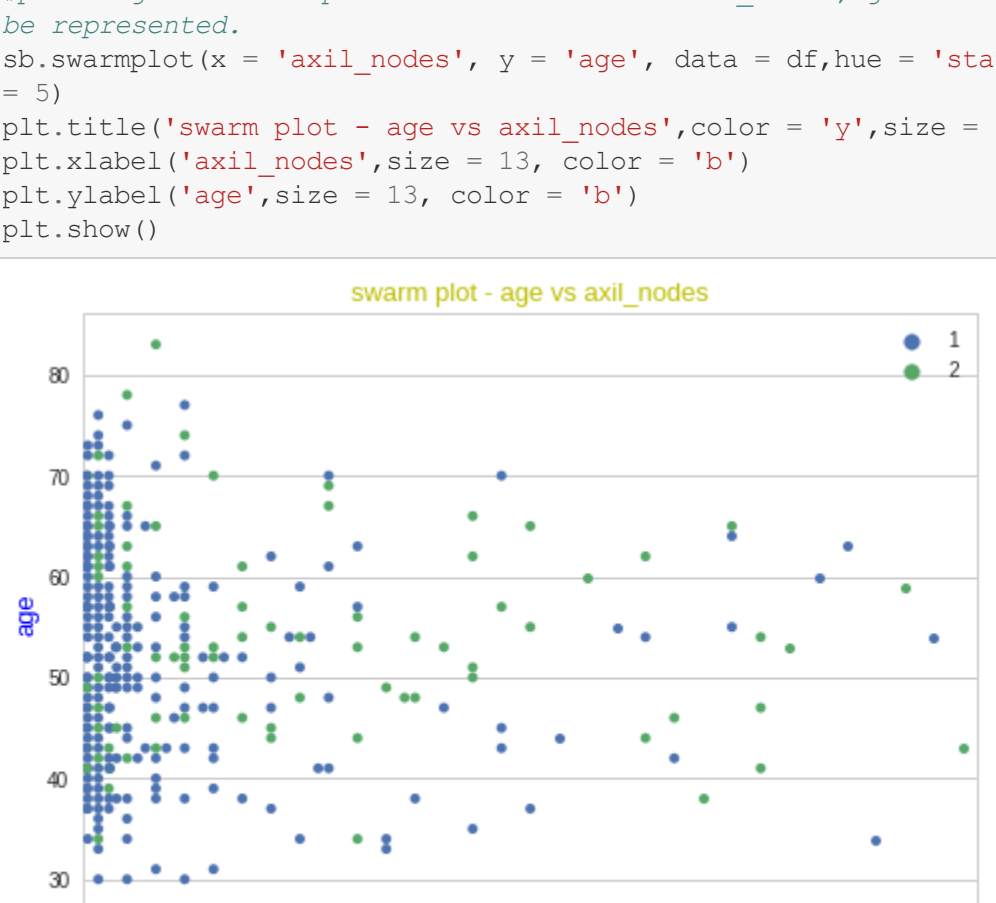
- **225 and 81 shows the count of patients who have survived more than 5 years and less than 5 years after the operation respectively.**

Bivariate Analysis

Scatterplot

In [140]:
```
#plotting the scatter plot
df.plot(kind = 'scatter',x = 'axil_nodes', y = 'age')
plt.title('scatter plot - age vs axil_nodes',color = 'y',size = '13')
plt.xlabel('axil_nodes',size = 13, color = 'b')
plt.ylabel('age',size = 13, color = 'b')
plt.show()
```

```
'c' argument looks like a single numeric RGB or RGBA sequence, which
should be avoided as value-mapping will have precedence in case its l
ength matches with 'x' & 'y'.  Please use a 2-D array with a single r
ow if you really want to specify the same RGB or RGBA value for all p
oints.
No handles with labels found to put in legend.
```

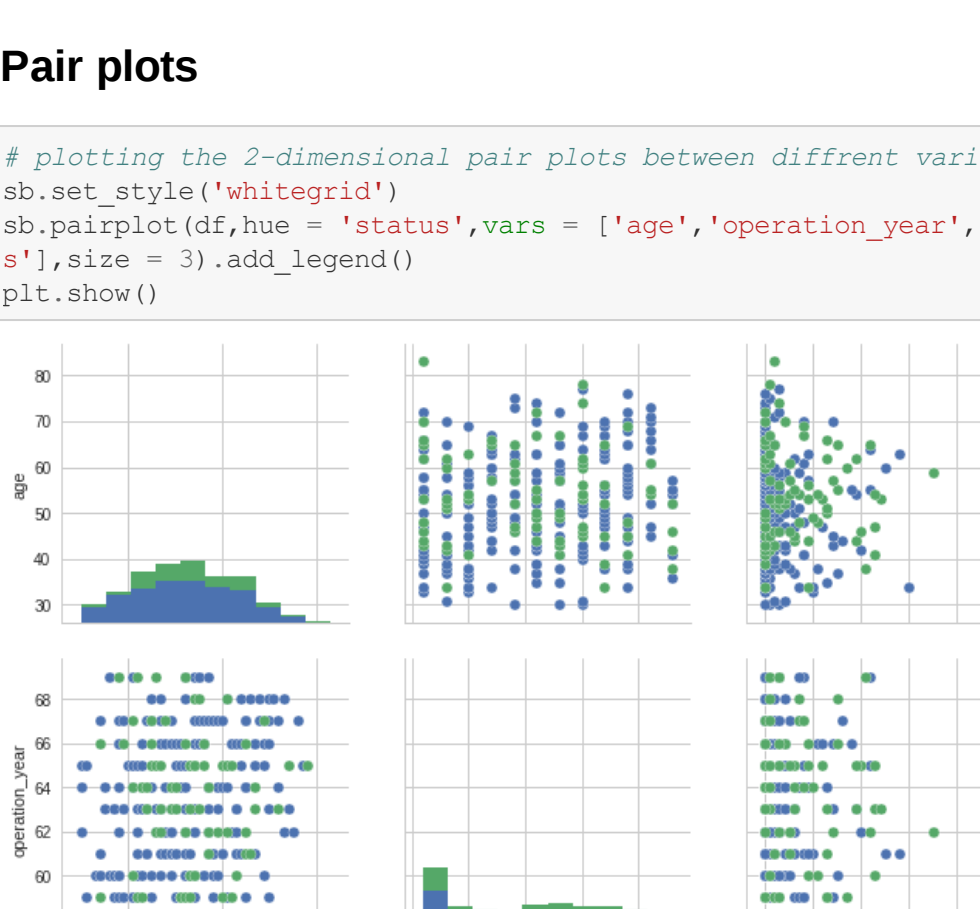


Observation:

- **From the above scatter plots we can infer that most the patients dont have axil_nodes (i.e axil_nodes =0)and very few patients have axil_nodes above 30**

Swarm plot

In [141]:
```
#plotting the swarm plot where data about axil_nodes,age and status can
be represented.
sb.swarmplot(x = 'axil_nodes', y = 'age', data = df,hue = 'status',size
= 5)
plt.title('swarm plot - age vs axil_nodes',color = 'y',size = '13')
plt.xlabel('axil_nodes',size = 13, color = 'b')
plt.ylabel('age',size = 13, color = 'b')
plt.show()
```

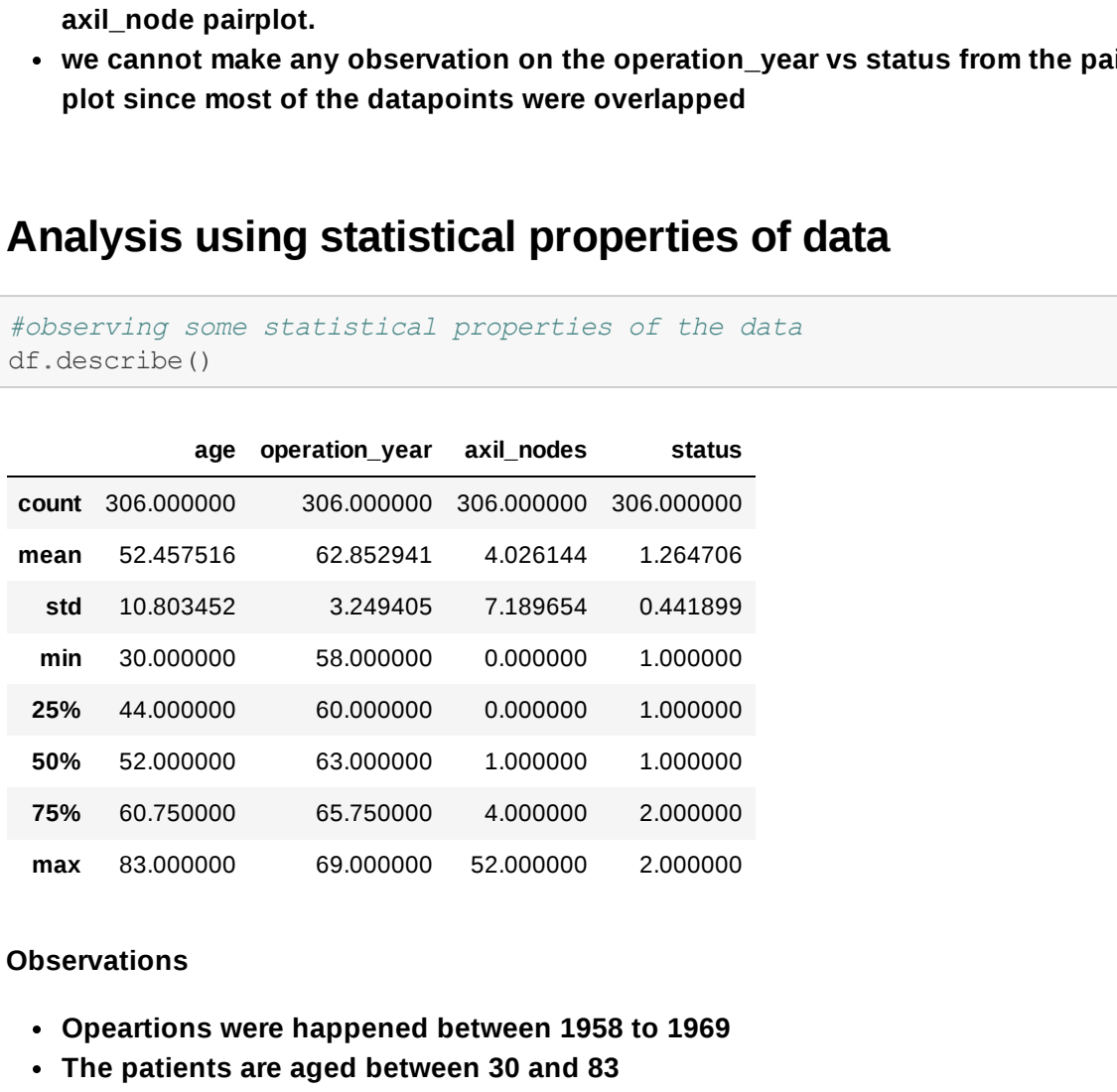


observation:

- **From the above plot, we can observe more number of patients have survived more than 5 years have 0 axil_nodes.**
- **More number of patients have axil_nodes < 5 and very less number of patients having more axil_nodes and in the plot.**
- **U can observe as the number of axial nodes increase the density of patients decreases gradually and patients having axial_nodes above 30 are very few**

Pair plots

In [121]:
```
# plotting the 2-dimensional pair plots between diffrent variable
sb.set_style('whitegrid')
sb.pairplot(df,hue = 'status',vars = ['age','operation_year','axil_node
s'],size = 3).add_legend()
plt.show()
```



Observations:

- **As observed in the swarm plot we can make same observations from the age vs axil_node pairplot.**
- **we cannot make any observation on the operation_year vs status from the pair plot since most of the datapoints were overlapped**

Analysis using statistical properties of data

In [122]:
```
#observing some statistical properties of the data
df.describe()
```

Out[122]:

	age	operation_year	axil_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

Observations

- **Opeartions were happened between 1958 to 1969**
- **The patients are aged between 30 and 83**
- **The mean of the status is 1.26 which little bit closer to 1 than 2 .So we can say that more than half of the people lived for more than 5 years after the operation**
- **There are patients with 0 axil_nodes**

In [0]:
```
#creating data frames for the data by splitting according to the status
df_gt5 = df.loc[df['status'] == 1]
df_lt5 =df.loc[df['status'] == 2]
```

In [125]:
```
df_gt5.describe()
```

Out[125]:

	age	operation_year	axil_nodes	status
count	225.000000	225.000000	225.000000	225.0
mean	52.017778	62.862222	2.791111	1.0
std	11.012154	3.222915	5.870318	0.0
min	30.000000	58.000000	0.000000	1.0
25%	43.000000	60.000000	0.000000	1.0
50%	52.000000	63.000000	0.000000	1.0
75%	60.000000	66.000000	3.000000	1.0
max	77.000000	69.000000	46.000000	1.0

Observations about the patients who lived more than 5 years:

- **The max age is 77**
- **Highest number of axil_nodes found is 46**
- **count of patients is 225 and mean of axil_nodes is 2.79**

In [126]:
```
df_lt5.describe()
```

Out[126]:

	age	operation_year	axil_nodes	status
count	81.000000	81.000000	81.000000	81.0
mean	53.679012	62.827160	7.456790	2.0
std	10.167137	3.342118	9.185654	0.0
min	34.000000	58.000000	0.000000	2.0
25%	46.000000	59.000000	1.000000	2.0
50%	53.000000	63.000000	4.000000	2.0
75%	61.000000	65.000000	11.000000	2.0
max	83.000000	69.000000	52.000000	2.0

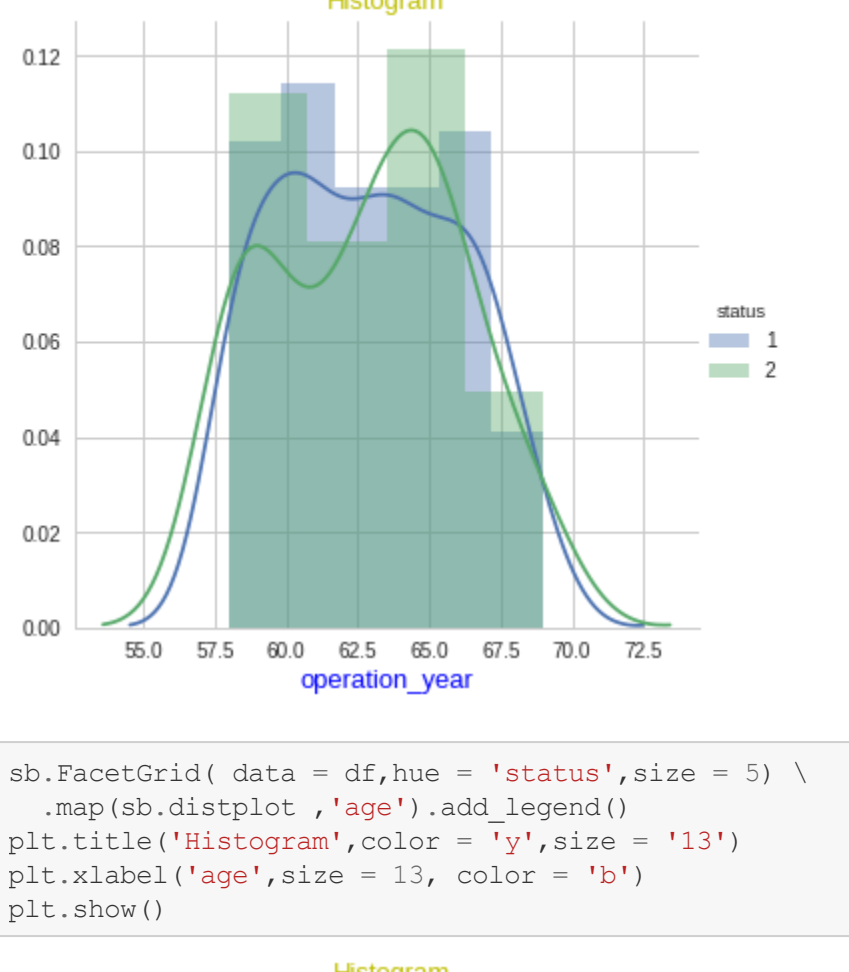
Observations

- **The mean of axil_nodes is 7.456 which is almost 2.5 times more than the axil_nodes we found in the patients who lived more than 5 years.**
- **so we can say that the patient having more than 46 axil_nodes died with in 5 years after the operation**
- **The patient having age more than 77 years have not survived more than 5 years after the operation**

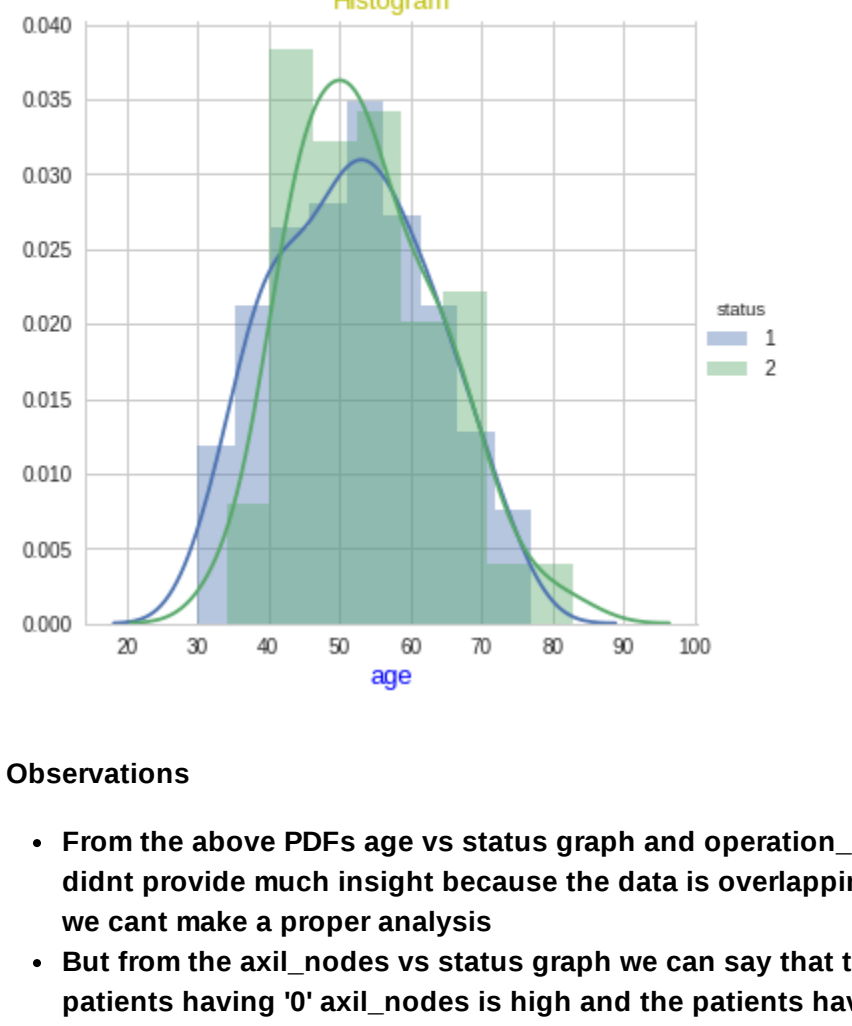
Univariate Analysis

Histograms and PDF

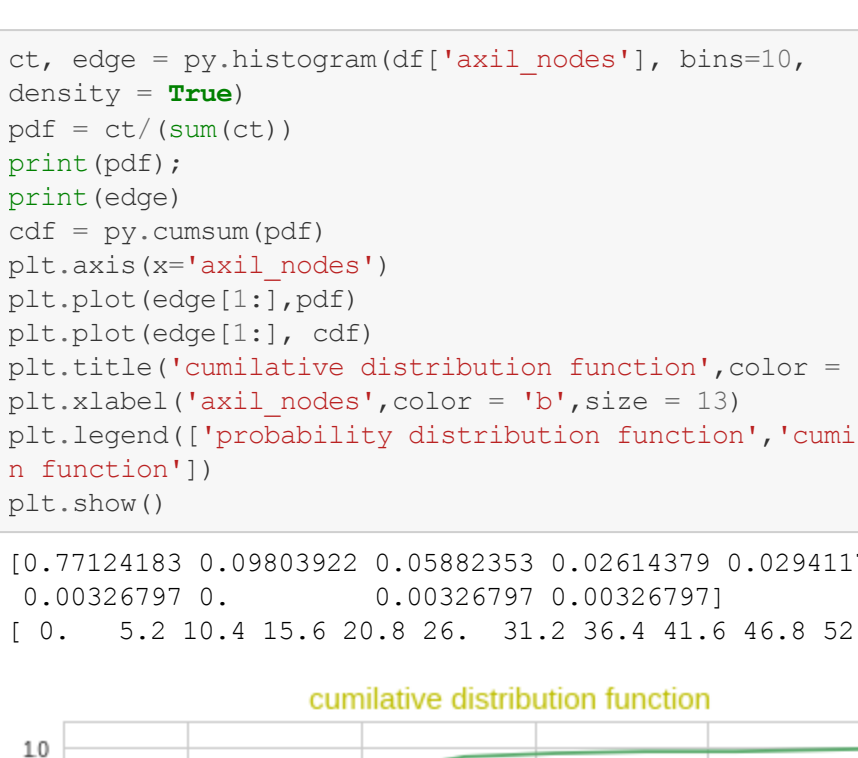
In [143]:
```
#Histogram plotting and the probability distribution function
sb.FacetGrid( data = df,hue = 'status',size = 5)\
  .map(sb.distplot ,'axil_nodes')\
  .add_legend()
plt.title('Histogram',color = 'y',size = '13')
plt.xlabel('axil_nodes',size = 13, color = 'b')
plt.show()
```



In [128]:
```
sb.FacetGrid( data = df,hue = 'status',size = 5) \
  .map(sb.distplot ,'operation_year').add_legend()
plt.title('Histogram',color = 'y',size = '13')
plt.xlabel('operation_year',size = 13, color = 'b')
plt.show()
```



In [129]:
```
sb.FacetGrid( data = df,hue = 'status',size = 5) \
  .map(sb.distplot ,'age').add_legend()
plt.title('Histogram',color = 'y',size = '13')
plt.xlabel('age',size = 13, color = 'b')
plt.show()
```



Observations

- **From the above PDFs age vs status graph and operation_year vs status graph didnt provide much insight because the data is overlapping in both plots.so we cant make a proper analysis**
- **But from the axil_nodes vs status graph we can say that the probability of patients having '0' axil_nodes is high and the patients having axil_nodes grater than 30 are very less**

CDF

In [130]:
```
ct, edge = py.histogram(df['axil_nodes'], bins=10,
density = True)
pdf = ct/(sum(ct))
print(pdf);
print(edge)
cdf = py.cumsum(pdf)
plt.axis(x='axil_nodes')
plt.plot(edge[1:],pdf)
plt.plot(edge[1:], cdf)
plt.title('cumilative distribution function',color = 'y',size = 15)
plt.xlabel('axil_nodes',color = 'b',size = 13)
plt.legend(['probability distribution function','cumilative distributio
n function'])
plt.show()
```

```
[0.77124183 0.09803922 0.05882353 0.02614379 0.02941176 0.00653595
 0.00326797 0.         0.00326797 0.00326797]
[ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
```

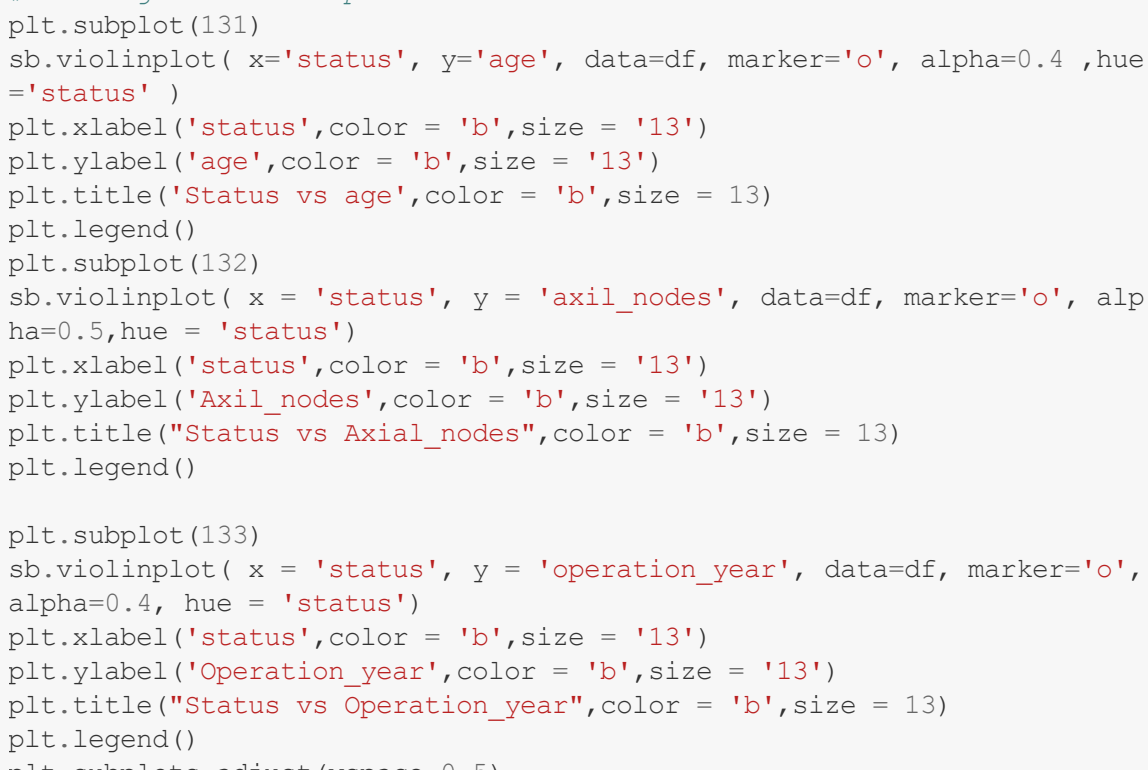
Observation

- **By observing the values we can say that the percentage of patients with '0' axil_nodes is 77 and with 52 axil_nodes is 0.3**
- **we can observe that the patients who are having axil_nodes greater than 30 are very few. U can clearly observe that the graph is closer to the axis as the count of axil_nodes are increasing**

Box plots and Strip plot

In [218]:
```
box=sb.boxplot(y='age', x='status', data=df, width=0.5,
                 palette="colorblind",hue = 'status')
box.legend(bbox_to_anchor=(1.2,0.75))
#plt.legend()
# adding stripplot to boxplot
box=sb.stripplot(y='age', x='status',
                   data=df,
                   jitter=True,
                   marker='o',
                   alpha=0.9,
                   color='black')
plt.xlabel('status',color = 'b',size = '13')
plt.ylabel('age',color = 'b',size = '13')
plt.title('Boxplot and Strip plot',color = 'y',size = 15)

plt.show()
```



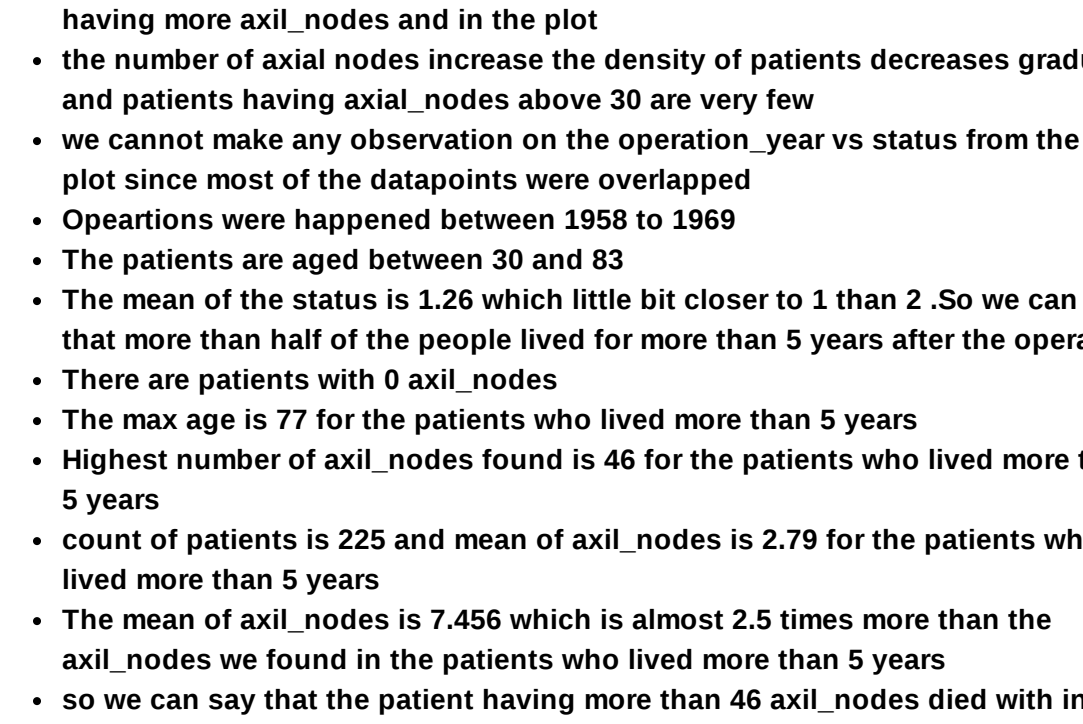
Observations

- **From the stripp plot we can say that more patients have lived for more than 5 years**
- **50% of patients have age less than 53 years**
- **75% of patients have age less than 62 years**

violin plots

In [178]:
```
#Plotting the violin plots
plt.subplot(131)
sb.violinplot( x='status', y='age', data=df, marker='o', alpha=0.4 ,hue
='status' )
plt.xlabel('status',color = 'b',size = '13')
plt.ylabel('age',color = 'b',size = '13')
plt.title('Status vs age',color = 'b',size = 13)
plt.legend()
plt.subplot(132)
sb.violinplot( x = 'status', y = 'axil_nodes', data=df, marker='o', alp
ha=0.5,hue = 'status')
plt.xlabel('status',color = 'b',size = '13')
plt.ylabel('Axil_nodes',color = 'b',size = '13')
plt.title("Status vs Axial_nodes",color = 'b',size = 13)
plt.legend()

plt.subplot(133)
sb.violinplot( x = 'status', y = 'operation_year', data=df, marker='o',
alpha=0.4, hue = 'status')
plt.xlabel('status',color = 'b',size = '13')
plt.ylabel('Operation_year',color = 'b',size = '13')
plt.title("Status vs Operation_year",color = 'b',size = 13)
plt.legend()
plt.subplots_adjust(wspace=0.5)
```



Summary of Observations

- **306 is the row count which indictes the number of patients and 4 is the column count which indicates the variable count**
- **225 and 81 shows the count of patients who have survived more than 5 years and less than 5 years after the operation respectively.**
- **From the scatter plots we can infer that most the patients dont have axil_nodes (i.e axil_nodes =0)and very few patients have axil_nodes above 30**
- **More number of patients have axil_nodes < 5 and very less number of patients having more axil_nodes and in the plot**
- **the number of axial nodes increase the density of patients decreases gradually and patients having axial_nodes above 30 are very few**
- **we cannot make any observation on the operation_year vs status from the pair plot since most of the datapoints were overlapped**
- **Opeartions were happened between 1958 to 1969**
- **The patients are aged between 30 and 83**
- **The mean of the status is 1.26 which little bit closer to 1 than 2 .So we can say that more than half of the people lived for more than 5 years after the operation**
- **There are patients with 0 axil_nodes**
- **The max age is 77 for the patients who lived more than 5 years**
- **Highest number of axil_nodes found is 46 for the patients who lived more than 5 years**
- **count of patients is 225 and mean of axil_nodes is 2.79 for the patients who lived more than 5 years**
- **The mean of axil_nodes is 7.456 which is almost 2.5 times more than the axil_nodes we found in the patients who lived more than 5 years**
- **so we can say that the patient having more than 46 axil_nodes died with in 5 years after the operation**
- **The patient having age more than 77 years have not survived more than 5 years after the operation**
- **From the PDFs age vs status graph and operation_year vs status graph didnt provide much insight because the data is overlapping in both plots.so we cant make a proper analysis**
- **By observing the values we can say that the percentage of patients with '0' axil_nodes is 77 and with 52 axil_nodes is 0.3**
- **From the CDF we can observe that the patients who are having axil_nodes greater than 30 are very few. U can clearly observe that the graph is closer to the axis as the count of axil_nodes are increasing beyond 30**
- **But from the axil_nodes vs status graph we can say that the probability of patients having '0' axil_nodes is high and the patients having axil_nodes grater than 30 are very less**
- **From the stripp plot we can say that more patients have lived for more than 5 years**
- **50% of patients have age less than 53 years**
- **75% of patients have age less than 62 years**

- **I have upload the same code to my github repository**
- **Here is the link : https://github.com/bharathpreetham/EDA-on-haberman-dataset/blob/master/haberman.ipynb**

In [ ]:

In [ ]:

In [ ]:

In [ ]: