Assignment : To perform EDA(with help of plotting techniques and staststical tools) on haberman dataset.

In [1]:
```
#importing python modules to accomplish EDA on Haberman dataset.

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

In [2]:
```
# to load dataset and renaming it to canc_patnt.
canc_patnt = pd.read_csv('D://Users//jalesh//Downloads/haberman.csv',header=None,
canc_patnt.tail(10)

#Assumption 1. class label 'survival_stat' has two categories 1 & 2 ,therefore,
# 1 would be: patients who survived more(>5 years) &,
# 2 would be: patiens who couldnt survive more(<5 years)

#Assumption 2: feature 'year_of_operation' would be considered as operation was c
```

Out[2]:

	age_when_operated	year_of_operation	Aux_lymph_nodes	survival_stat
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

In [18]:
```
# to check no of datapoints,features & consice summary of canc_patnt dataset.
canc_patnt.shape
canc_patnt.columns
canc_patnt.describe()

#Observation:
#1. average age of pateients are ~52.4
#2. high deviation(scatterness) is seen in age_when_operted feature ~ 10.8
#3. Aux_lymph_nodes feature is spreaded more around its mean  ~ 7.18
#4. Minimum and maximum age was 30 & 83 years respectively.
#5. 75% os the Aux_lymph_nodes feature has value <= 4.
#6. total no of operations performed were 306
#7. 50% of operations performed between age ~ 44 and 61 years
```

Out[18]:

	age_when_operated	year_of_operation	Aux_lymph_nodes	survival_stat
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

Objective : Statistical tools application in performaing EDA on canc_patnt dataframe

In [4]:
```
canc_patnt['survival_stat'].value_counts()

#Observation:
#1.class label survival_stat is imbalaced as it has hetregenous count of 1(who su
#2.81 patients operated upon and who couldnt survived more(2)
#3.225 people were operated upon and could survive more.(1)
#4.306 people were operated upon in totality.
```

Out[4]:
```
1    225
2     81
Name: survival_stat, dtype: int64
```

In [9]:
```
for i in (canc_patnt['age_when_operated'].mean(),canc_patnt['age_when_operated'].
          canc_patnt['Aux_lymph_nodes'].mean(),canc_patnt['Aux_lymph_nodes'].std(
    print(i, end=' ')


#observation: there's high amount of deviation in both features 'age_when_operate
```

```
52.45751633986928 10.80345234930328 4.026143790849673 7.189653506248565
```

In [10]:
```
canc_patnt['year_of_operation'].value_counts()

#Observation: 1958 was the year when maximum no of cancer petients were operatoed
#             1969 was the year when operation done was minimum.
```

Out[10]:
```
58    36
64    31
63    30
66    28
65    28
60    28
59    27
61    26
67    25
62    23
68    13
69    11
Name: year_of_operation, dtype: int64
```

In [16]:
```
for i in (canc_patnt.groupby('survival_stat').max(), canc_patnt.groupby('survival_
    print(pd.DataFrame(i))
    print('*' * 100)



#Observation: 1. max age of person who could(1) and couldnt survive(2) more is 77
#             2. min age of person who could(1) and couldnt survive(2) more is 30
```

```
               age_when_operated  year_of_operation  Aux_lymph_nodes
survival_stat
1                             77                 69               46
2                             83                 69               52
****************************************************************************
********************
               age_when_operated  year_of_operation  Aux_lymph_nodes
survival_stat
1                             30                 58                0
2                             34                 58                0
****************************************************************************
********************
```

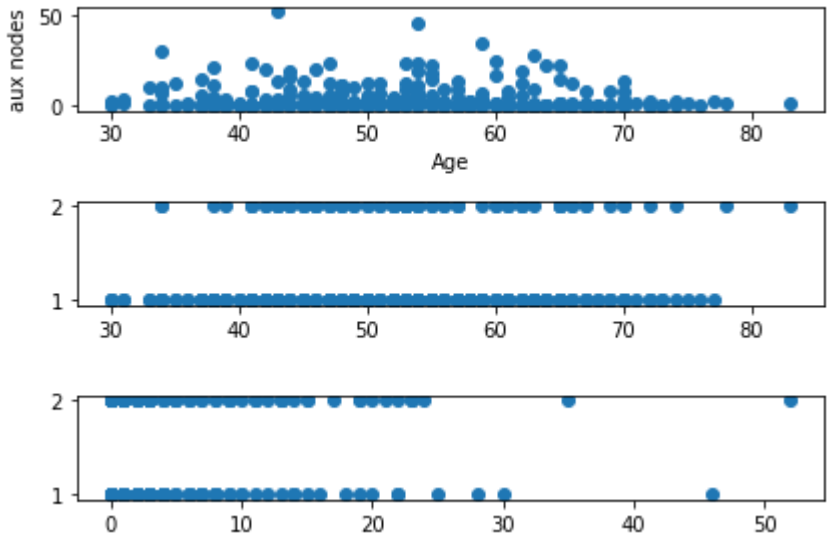
In [17]:
```
canc_patnt.groupby('survival_stat').std()
#observation:
#1. the average person's age in who survived more and less list are ~52 and ~53 r
#2. the avgerage person's age in who survived more had lower no of Aux_lymph_node
```

Out[17]:

survival_stat	age_when_operated	year_of_operation	Aux_lymph_nodes
1	11.012154	3.222915	5.870318
2	10.167137	3.342118	9.185654

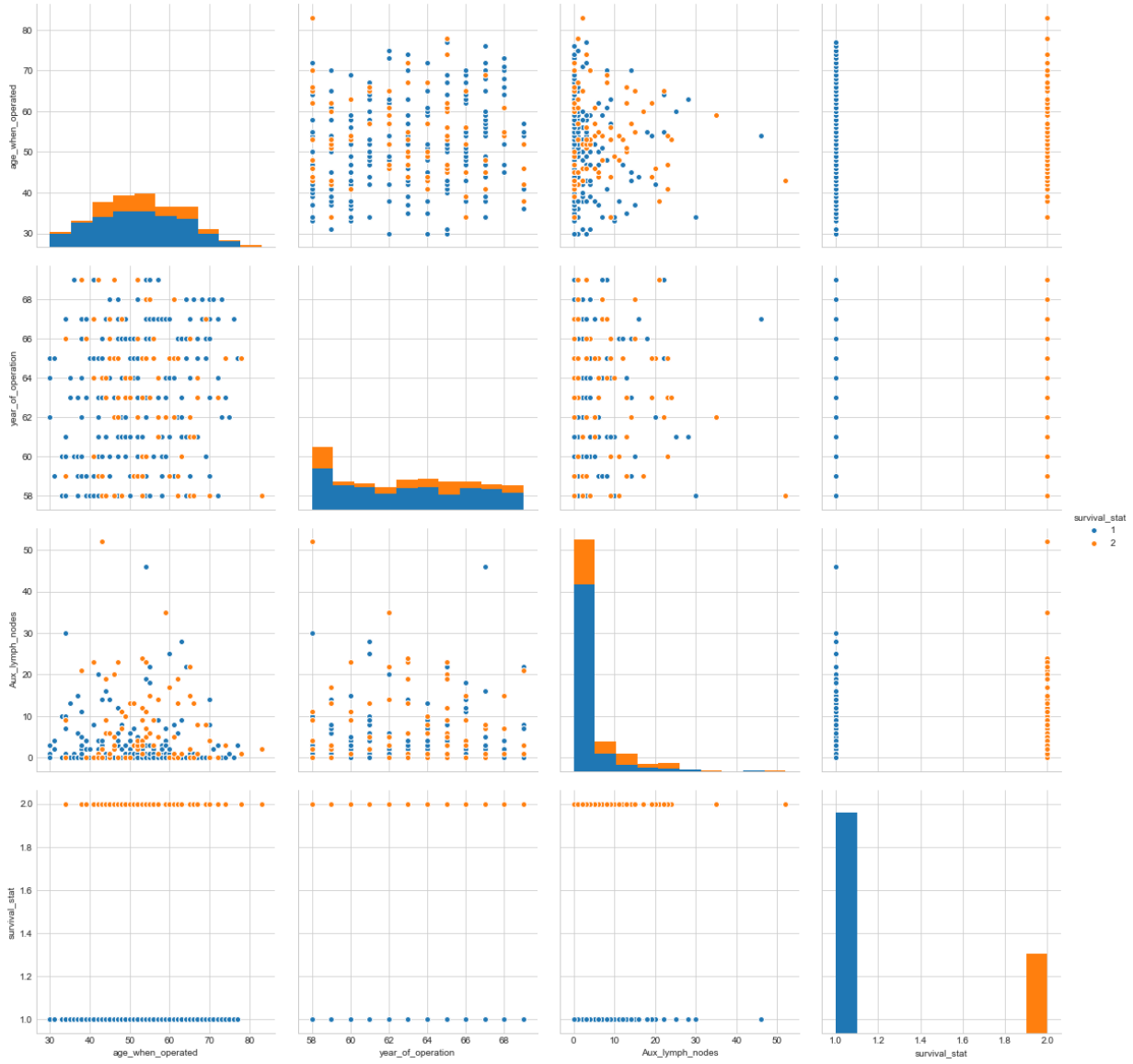
Objective:graphical analysis on canc_patnt dataframe using various plotting method

In [26]:
```
#scatter plot

fig = plt.Figure(figsize=(14,14))

plt.subplot(3,1,1)
plt.plot('age_when_operated', 'Aux_lymph_nodes', 'o',data=canc_patnt )
plt.xlabel('Age')
plt.ylabel('aux nodes')

plt.subplot(3,1,2)
plt.plot('age_when_operated', 'survival_stat', 'o',data=canc_patnt)

plt.subplot(3,1,3)
plt.plot('Aux_lymph_nodes', 'survival_stat', 'o', data=canc_patnt)

plt.tight_layout()
```



In [27]:
```
#pairplot
sns.set_style('whitegrid')
sns.pairplot(canc_patnt, hue='survival_stat', size=4)


#observation:
#1.Aux_lynph_nodes can be a good feature in segregeating the class label
```

Out[27]: <seaborn.axisgrid.PairGrid at 0x8c02dd0>



In [28]:
```
canc_patnt.corr()

#Observation: Aux_lymph_nodes feautre has highest correlation wrt class label sur
```

Out[28]:

	age_when_operated	year_of_operation	Aux_lymph_nodes	survival_stat
age_when_operated	1.000000	0.089529	-0.063176	0.067950
year_of_operation	0.089529	1.000000	-0.003764	-0.004768
Aux_lymph_nodes	-0.063176	-0.003764	1.000000	0.286768
survival_stat	0.067950	-0.004768	0.286768	1.000000

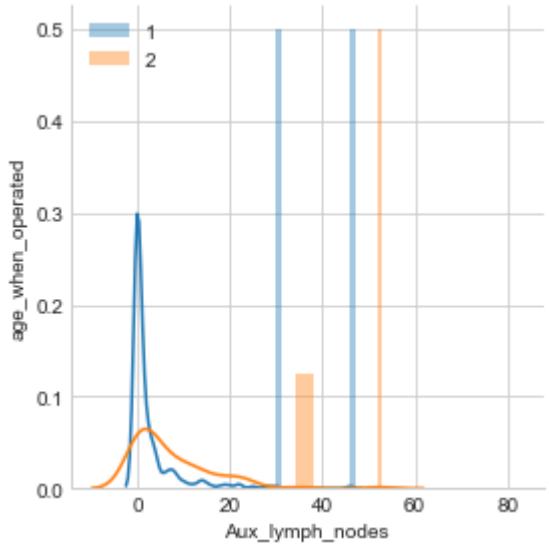
In [29]:
```
sum(canc_patnt[(canc_patnt['Aux_lymph_nodes'] <=10) & (canc_patnt['survival_stat'

#observation:
#1.through pairplot ,patients having aux_lymph_nodes ~ <10 survived more years ha
```

Out[29]: 208

In [30]:
```
#plotting univariate distribution plots in one canvas:

sns.set_style("whitegrid")
g = sns.FacetGrid(canc_patnt, hue="survival_stat", size=4)
g.map(sns.distplot, 'Aux_lymph_nodes', 'age_when_operated')
plt.legend()
#observation:patents couldnt suvive more had more wider of spread of lymphs as co
```

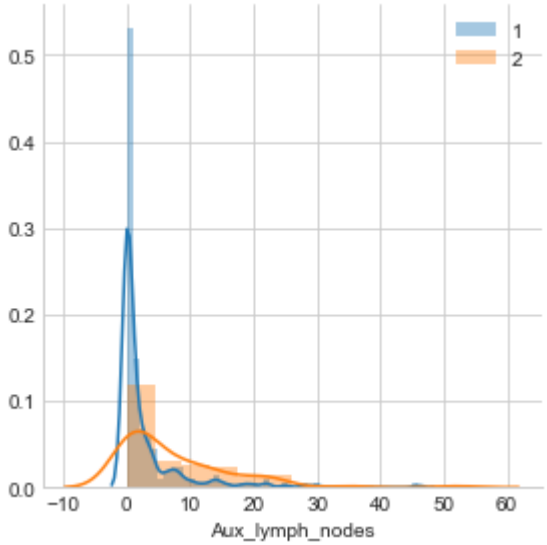
```
E:\anaconda\lib\site-packages\matplotlib\axes\_axes.py:6201: RuntimeWarning: in
valid value encountered in true_divide
  m = (m.astype(float) / db) / m.sum()
```

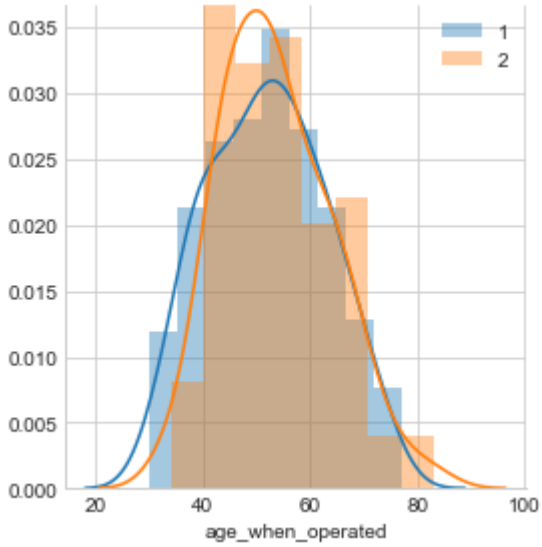
Out[30]: <matplotlib.legend.Legend at 0xc18dab0>



In [32]:
```
#comparing two univariate plots:

sns.set_style("whitegrid")

g = sns.FacetGrid(canc_patnt, hue="survival_stat", size=4)
g.map(sns.distplot, 'Aux_lymph_nodes')
plt.legend()

g = sns.FacetGrid(canc_patnt, hue="survival_stat", size=4)
g.map(sns.distplot, 'age_when_operated')
plt.legend()

#observation: comparing two univarete distribution plots , distribution graph of 
#             'age_when_operated' as point of intersection of two PDFs are lesser
```

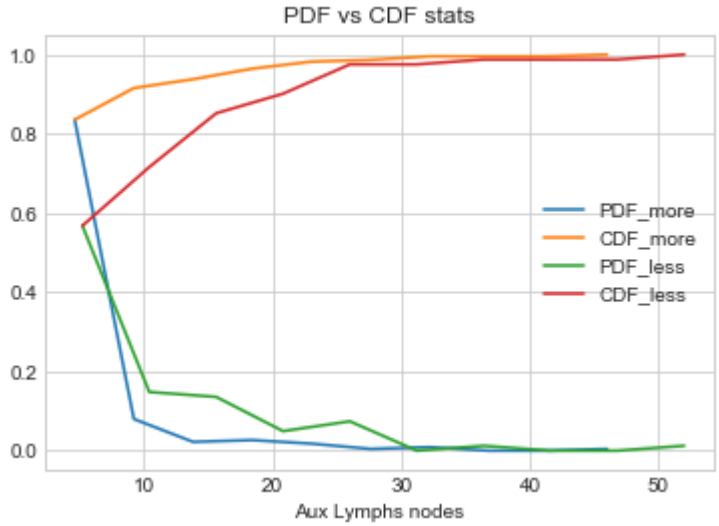
Out[32]: <matplotlib.legend.Legend at 0xd546430>





In [34]:
```
# to plot PDF vs CDF in same figure:

fig = plt.Figure(figsize=(10,10), dpi=400)

#creating dataframe of those patients who survided more aftre having operated
survived_more = canc_patnt[canc_patnt['survival_stat'] == 1]


counts, bin_edges = np.histogram(survived_more['Aux_lymph_nodes'], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)

# plotting 1st PDF vs CDF:

plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)
plt.title('PDF vs CDF stats')
plt.xlabel('Aux Lymphs nodes')


#creating dataframe of those patients who survided less after having operated

survived_less = canc_patnt[canc_patnt['survival_stat'] == 2]

counts, bin_edges = np.histogram(survived_less['Aux_lymph_nodes'], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)

#plotting 2nd PDF vs CDF
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)
plt.title('PDF vs CDF stats')
plt.xlabel('Aux Lymphs nodes')
plt.legend(['PDF_more', 'CDF_more','PDF_less', 'CDF_less'])


#observation: around 81% of patient who had 'Aux_lymph_node' ~ <=10 have survived
#           : ~85% of patients who had 'Aux_lymph_node' ~ >=15 have survived less
```

Out[34]: <matplotlib.legend.Legend at 0xea42db0>



In [38]:
```
survived_more.describe()
#Observation: average lymphs ~3 with +-6 (SD)
```

Out[38]:

	age_when_operated	year_of_operation	Aux_lymph_nodes	survival_stat
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

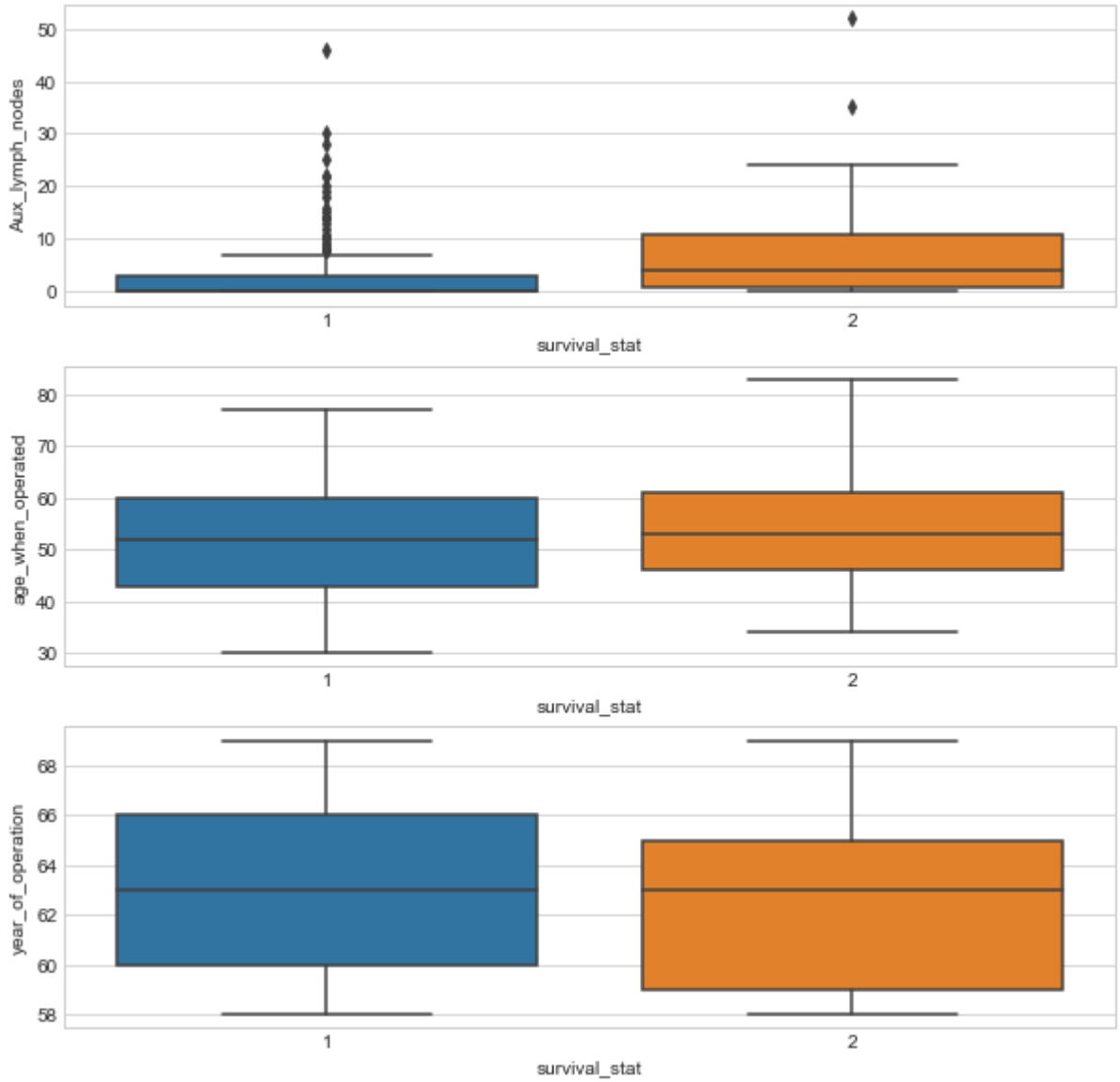
In [39]:
```
survived_less.describe()
#Observation: average lymphs ~7.5 with +-9 (SD)
```

Out[39]:

	age_when_operated	year_of_operation	Aux_lymph_nodes	survival_stat
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

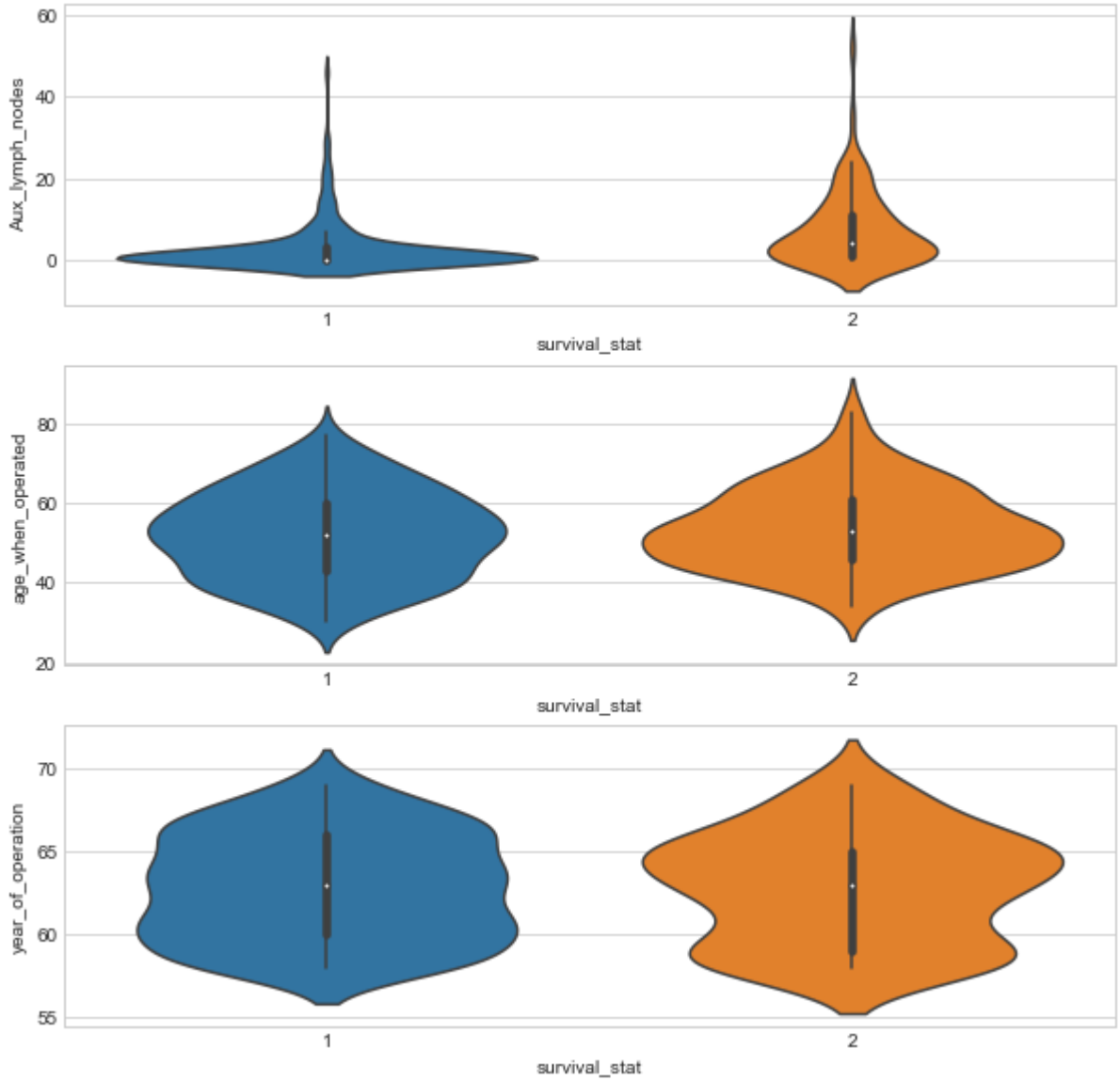
In [35]:
```
#boxplot to get the value of 25/50 75 percentile value

fig = plt.figure(figsize=(10,10))
plt.subplot(3,1,1)
sns.boxplot(x='survival_stat', y='Aux_lymph_nodes', data=canc_patnt)

plt.subplot(3,1,2)
sns.boxplot(x='survival_stat', y='age_when_operated', data=canc_patnt)

plt.subplot(3,1,3)
sns.boxplot(x='survival_stat', y='year_of_operation', data=canc_patnt)

#observation:
#1. min and 25% quantile have same value
#2. all the patients had 0 Aux_lymph_nodes survived more(1) after having operated
#3. those who were operated before 1960 , couldnt live more (2) after having opera
```

Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0xeba0030>



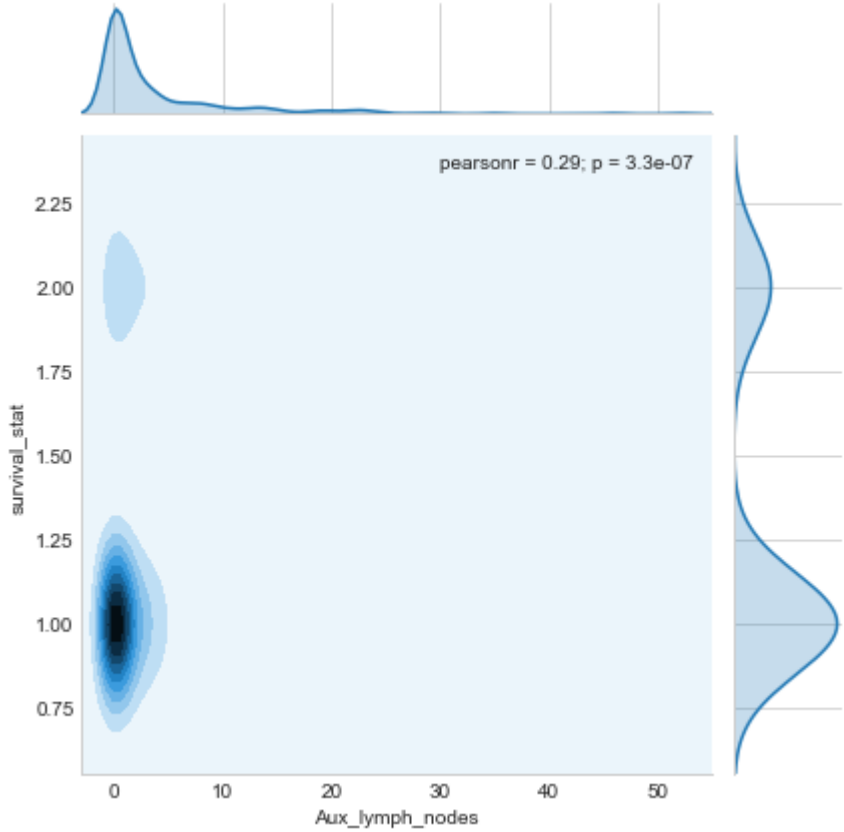
In [36]:
```
#violin plot
fig = plt.figure(figsize=(10,10))
plt.subplot(3,1,1)
sns.violinplot(x='survival_stat', y='Aux_lymph_nodes', data=canc_patnt)

plt.subplot(3,1,2)
sns.violinplot(x='survival_stat', y='age_when_operated', data=canc_patnt)

plt.subplot(3,1,3)
sns.violinplot(x='survival_stat', y='year_of_operation', data=canc_patnt)
```

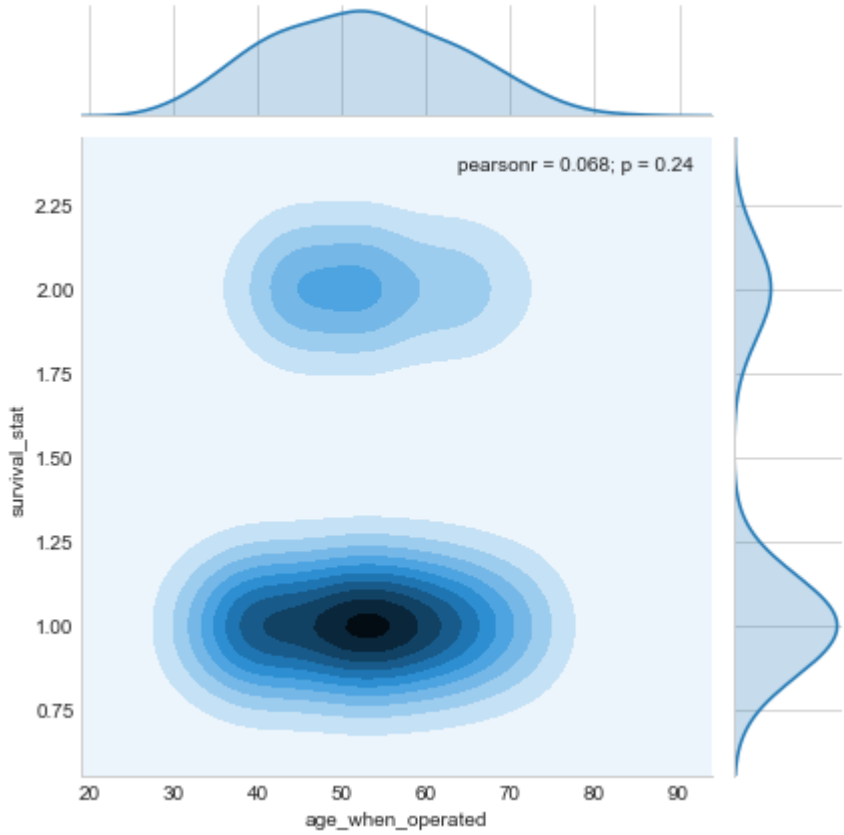
Out[36]: <matplotlib.axes._subplots.AxesSubplot at 0xeca58d0>

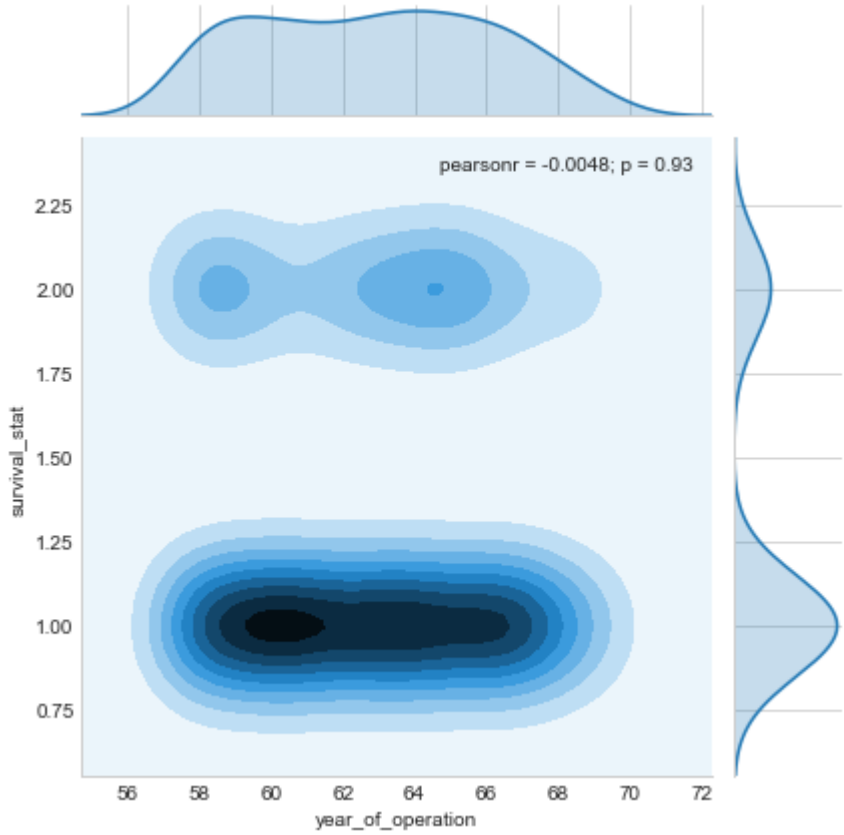


In [37]:
```
#bivariate jointplot:
fig = plt.Figure(figsize=(4,10), dpi=200)

sns.jointplot(y='survival_stat', x='Aux_lymph_nodes', data=canc_patnt, kind='kde'
sns.jointplot(y='survival_stat', x='age_when_operated', data=canc_patnt, kind='kd
sns.jointplot(y='survival_stat', x='year_of_operation', data=canc_patnt, kind='kd

#observation:
#1. age between 40 to 70 and lived longer
#2. ~1960 was the year when most no of operations were performed.
```

Out[37]: <seaborn.axisgrid.JointGrid at 0xd8810f0>




pearsonr = 0.068; p = 0.24
survival_stat
2.25
2.00
1.75
1.50
1.25
1.00
0.75
20
30
40
50
60
70
80
90
age_when_operated


pearsonr = -0.0048; p = 0.93
survival_stat
2.25
2.00
1.75
1.50
1.25
1.00
0.75
56
58
60
62
64
66
68
70
72
year_of_operation

In [ ]: