Haberman's Survival Data Set

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.

In [122]:

```
# importing libraries for the given data set.
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
# 1.)loading the data frame set (df) by using pandas.
haber=pd.read_csv('haber.csv')  # data set from kaggle.(Source:
https://www.kaggle.com/gilsousa/habermans-survival-data-set/data)
```

Infromation about the dataset:

In this haberman's data set there are four columns. First three columns represents information regarding patients(age,year,node) and fourth column is about survial status of the patients.

In [116]:

```
#2.)
print(haber.shape)#The shape determines 306 rows and 4 columns.
print(haber.size)# It represents the size of the dataset.
print(haber.ndim)#It represents the dimensions of the dataset.
```

```
(306, 4)
1224
2
```

Objective

The main objective of this dataset is to identify the parameters which can help us to determine if the patient survives or not .Hence we can say it has two classes i.e. SURVIVAL(status - 1) and NOTSURVIVAL(status - 2).

In [123]:

```
# ''''value_counts(): { It determines classes of the status 1 & 2.}
# In this data set there are 2 classes for status. 1.) survival -225 2.) Nonsurvival -81

haber['status'].value_counts()
```

Out[123]:

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

In [7]:

```
# It determines the no of columns in the dataset.
haber.columns
```

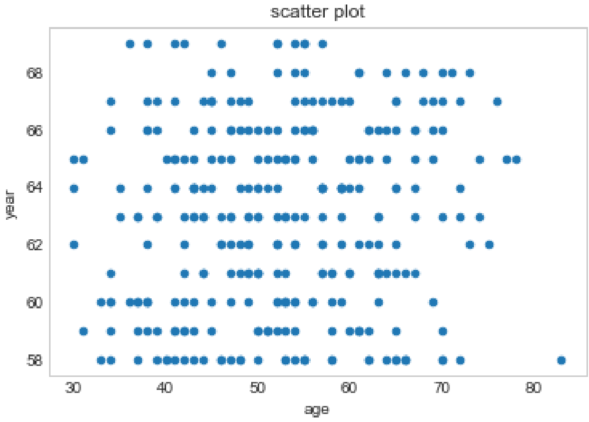
Out[7]:

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

2D- scatter plot

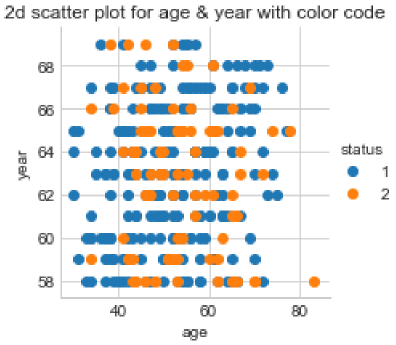
In [220]:

```
# 2d scatter plot  (scattering )describing kind='scatter' for age & year.

haber.plot(kind='scatter',x='age',y='year')
plt.grid()
plt.title('scatter plot')
plt.show()
```



In [237]:

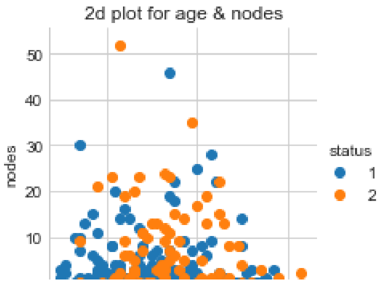
```
# 2d plt for age and year with color.using seaborn sns command.hue is color or shade.
# size or height represents figure size.
# FacetGrid used to create a template .map & plt.scatter

sns.set_style('whitegrid')
sns.FacetGrid(haber,hue='status',size=3)\
    .map(plt.scatter,'age','year')\
    .add_legend()
plt.title('2d scatter plot for age & year with color code')
plt.show()
```

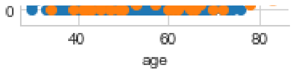


observations: Here by using status 1 & 2 for age we can determine survival & not survival data set They are combined using Age and year features. so, the classifiaction between both is not possible.

In [222]:

```
# 2d- for age & node
sns.set_style('whitegrid')
sns.FacetGrid(haber,hue='status',size=3)\
    .map(plt.scatter,'age','nodes')\
    .add_legend()
plt.title('2d plot for age & nodes')
plt.show()
```



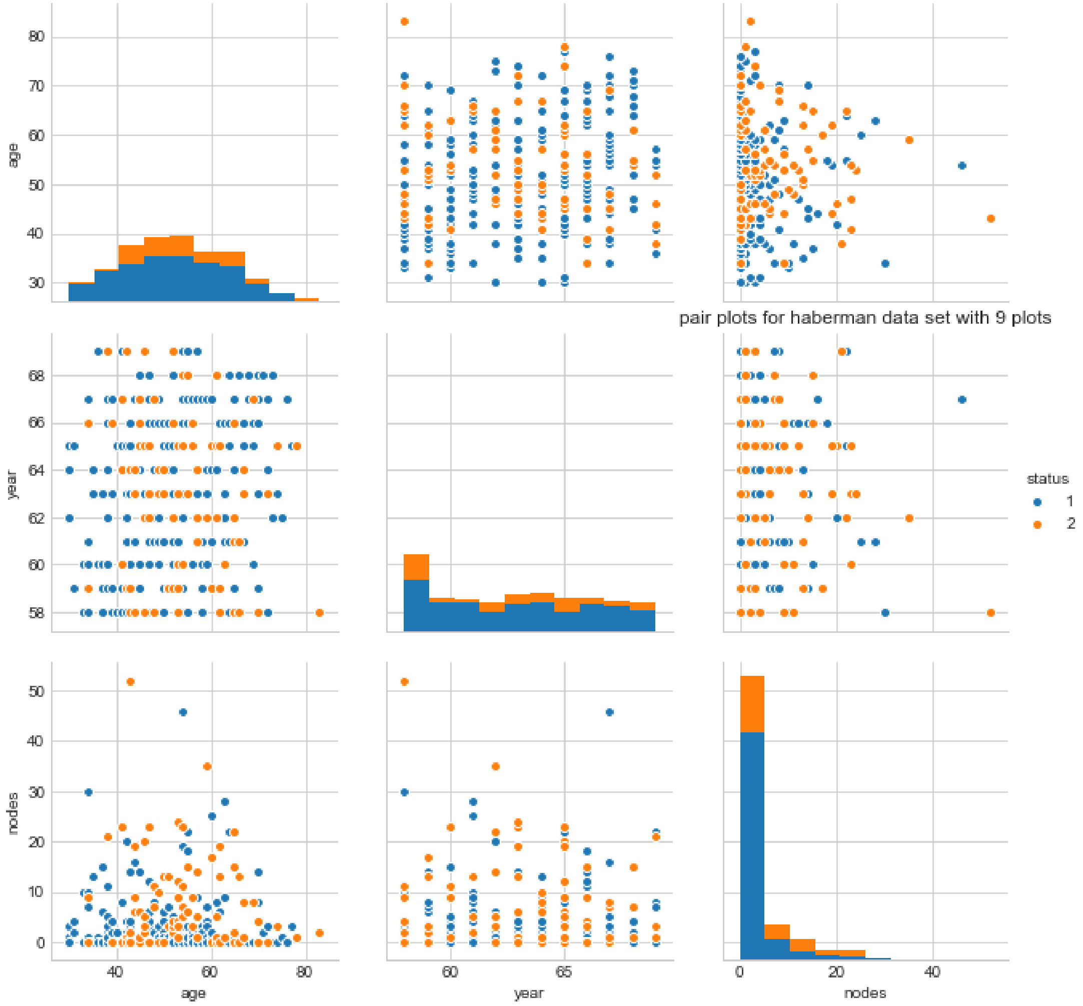


observations: Here also same status 1 & 2 are combined using Age and nodes features. so, the classifiaction between both is not possible.

pair plots

In [225]:

```
# 3d & upto nd scatter plots are genrally represented in pairplots
# represented using pair plots in 3c2 ways. using variable command vars

plt.close();
sns.set_style('whitegrid')
sns.pairplot(haber,hue='status',vars=['age','year','nodes'],size=3)
plt.title('pair plots for haberman data set with 9 plots')
plt.show()
```



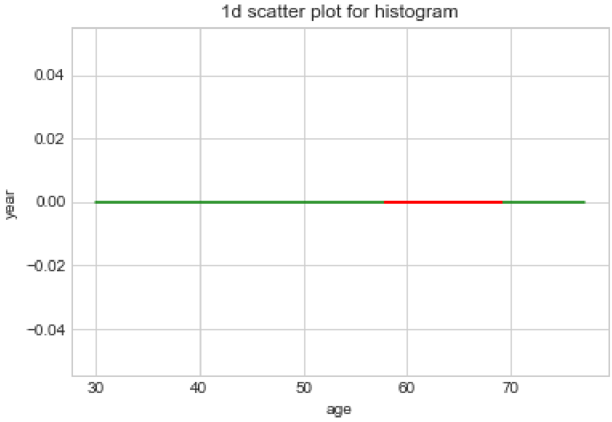
observations: As we are unable to classify which is the most useful feature because of imbalanced data set.

Histogram pdf & cdf

Introduction: Probality Density Function (PDF) is the probabilty that the variable takes a value x. (smoothed version of the histogram) Kernel Density Estimate (KDE) is the way to estimate the PDF. The area under the KDE curve is 1. Here the height of the bar denotes the percentage of data points under the corresponding group

In [226]:

```
# using 1d -scatter plot for histogram pdf

habersurvive=haber.loc[haber['status']==1];
habernotsurvive=haber.loc[haber['status']==2];

plt.plot(habersurvive['age'],np.zeros_like(habersurvive['age']),'g')
plt.plot(habernotsurvive['year'],np.zeros_like(habernotsurvive['year']),'r')
plt.xlabel('age')
plt.ylabel('year')
plt.title('1d scatter plot for histogram')
plt.show()
```



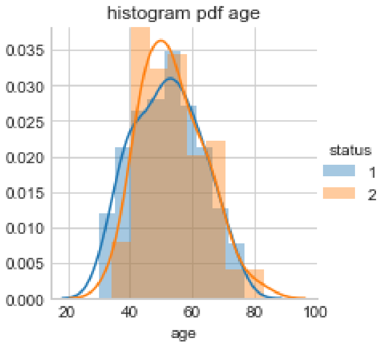
observations: we cannot counts the no of points using 1D scatter plot

Distrubution plots(sns.distplot)for histogram pdfs

In [227]:

```
#pdf for age
sns.set_style('whitegrid')
sns.FacetGrid(haber,hue='status',size=3)\
    .map(sns.distplot,'age')\
    .add_legend()
plt.title('histogram pdf age')
plt.show()
```

```
C:\Users\Rohith\python\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed'
kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\Rohith\python\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed'
kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

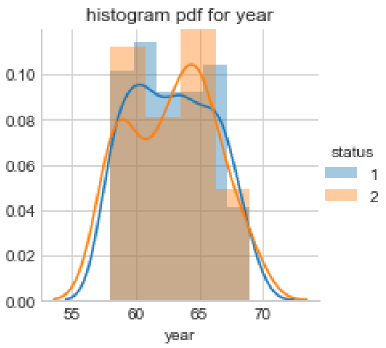


observation: This histogram is overlapping each other, but still we can say that people within range of 40-60 are nonsurvival& People less than age 40 are more likely to survive

In [228]:

```
#pdf for year
sns.set_style('whitegrid')
sns.FacetGrid(haber,hue='status',size=3)\
    .map(sns.distplot,'year')\
    .add_legend()
plt.title('histogram pdf for year')
plt.show()
```

```
C:\Users\Rohith\python\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed'
kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\Rohith\python\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed'
kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

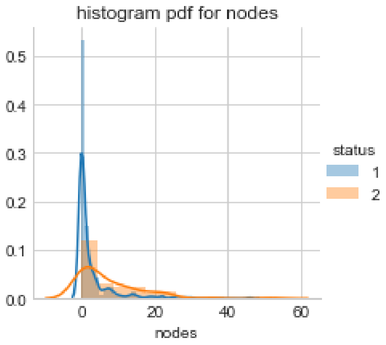


observation: more number of patients in year 60 and 65 are nonsurvival

In [229]:

```
#pdf for nodes
sns.set_style('whitegrid')
sns.FacetGrid(haber,hue='status',size=3)\
    .map(sns.distplot,'nodes')\
    .add_legend()
plt.title('histogram pdf for nodes')
plt.show()
```

```
C:\Users\Rohith\python\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed'
kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
C:\Users\Rohith\python\lib\site-packages\matplotlib\axes\_axes.py:6462: UserWarning: The 'normed'
kwarg is deprecated, and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



observation: Patients having less than 0 axil nodes are more likely to survive

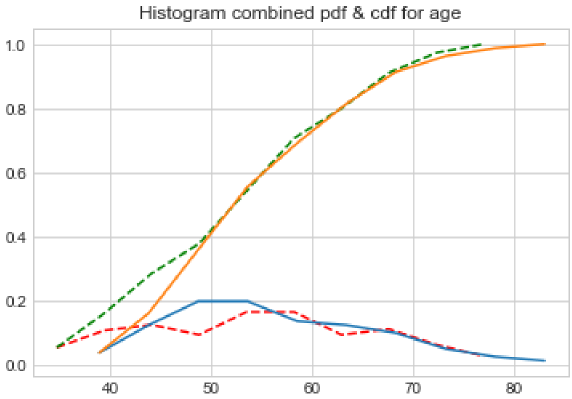
Histogram cdf

In [230]:

```
# histogram cdf for survival and not survival patients

 #  1.) survival & not survival for age
#Survival
counts,bin_edges=np.histogram(habersurvive['age'],bins=10,density=True)
pdf=counts/(sum(counts));
print(pdf)
print(bin_edges)
cdf=np.cumsum(pdf)#cummilative sum from linear algebra
plt.plot(bin_edges[1:],pdf,'r--')
plt.plot(bin_edges[1:],cdf,'g--')


#notSurvival
counts,bin_edges=np.histogram(habernotsurvive['age'],bins=10,density=True)
pdf=counts/(sum(counts));
print(pdf)
print(bin_edges)
```

```
cdf=np.cumsum(pdf)#cummilative sum from linear algebra
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.title('Histogram combined pdf & cdf for age')
plt.show()
```

```
[0.05333333 0.10666667 0.12444444 0.09333333 0.16444444 0.16444444
 0.09333333 0.11111111 0.06222222 0.02666667]
[30.   34.7 39.4 44.1 48.8 53.5 58.2 62.9 67.6 72.3 77. ]
[0.03703704 0.12345679 0.19753086 0.19753086 0.13580247 0.12345679
 0.09876543 0.04938272 0.02469136 0.01234568]
[34.   38.9 43.8 48.7 53.6 58.5 63.4 68.3 73.2 78.1 83. ]
```

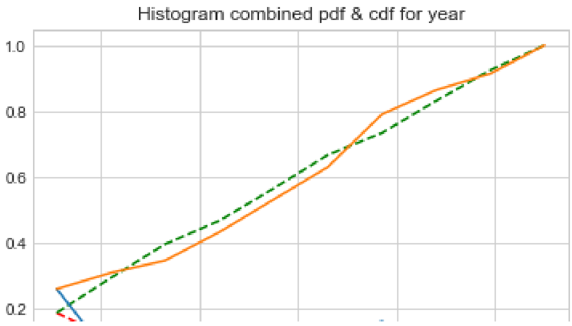


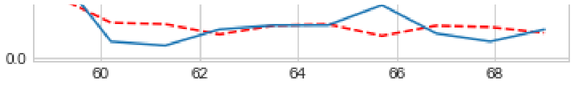
Observations: 1.)The pdf are almost overlapping hence we cannot make model basesd on age.Although there are more number of patients survied.

In [231]:

```
 #  1.) survival & not survival for year
#Survival
counts,bin_edges=np.histogram(habersurvive['year'],bins=10,density=True)
pdf=counts/(sum(counts));
print(pdf)
print(bin_edges)
cdf=np.cumsum(pdf)#cummilative sum from linear algebra
plt.plot(bin_edges[1:],pdf,'r--')
plt.plot(bin_edges[1:],cdf,'g--')


#notSurvival
counts,bin_edges=np.histogram(habernotsurvive['year'],bins=10,density=True)
pdf=counts/(sum(counts));
print(pdf)
print(bin_edges)
cdf=np.cumsum(pdf)#cummilative sum from linear algebra
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.title('Histogram combined pdf & cdf for year')
plt.show()
```

```
[0.18666667 0.10666667 0.10222222 0.07111111 0.09777778 0.10222222
 0.06666667 0.09777778 0.09333333 0.07555556]
[58.   59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
[0.25925926 0.04938272 0.03703704 0.08641975 0.09876543 0.09876543
 0.16049383 0.07407407 0.04938272 0.08641975]
[58.   59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
```



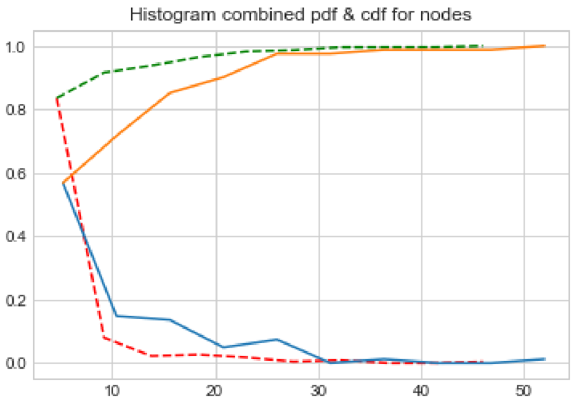


Observations 1.)Here also both pdfs overlapping and also it is not as good as age pdfs .so taking feature of year is also ruled out.

In [232]:

```
 #  1.) survival & not survival for nodes
#Survival
counts,bin_edges=np.histogram(habersurvive['nodes'],bins=10,density=True)
pdf=counts/(sum(counts));
print(pdf)
print(bin_edges)
cdf=np.cumsum(pdf)#cummilative sum from linear algebra
plt.plot(bin_edges[1:],pdf,'r--')
plt.plot(bin_edges[1:],cdf,'g--')


#notSurvival
counts,bin_edges=np.histogram(habernotsurvive['nodes'],bins=10,density=True)
pdf=counts/(sum(counts));
print(pdf)
print(bin_edges)
cdf=np.cumsum(pdf)#cummilative sum from linear algebra
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)
plt.title('Histogram combined pdf & cdf for nodes')
plt.show()
```

```
[0.83555556 0.08       0.02222222 0.02666667 0.01777778 0.00444444
 0.00888889 0.         0.         0.00444444]
[ 0.   4.6  9.2 13.8 18.4 23.  27.6 32.2 36.8 41.4 46. ]
[0.56790123 0.14814815 0.13580247 0.04938272 0.07407407 0.
 0.01234568 0.         0.         0.01234568]
[ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
```

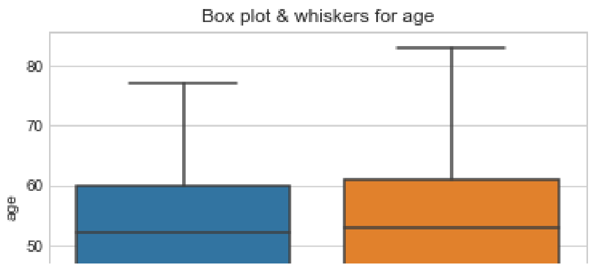

Histogram combined pdf & cdf for nodes

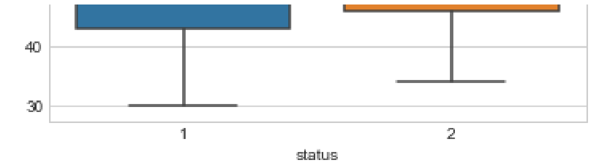
Observations: 1.)if number of nodes are more 47 than No one survives 2.)If number of nodes are less than 8 than survive. 3.)Generally in this data set no of survival status is more

Box plot & whiskers

In [233]:

```
#box plot are used to represent visually as a pdf of histogram
sns.boxplot(x='status',y='age',data=haber)
plt.title('Box plot & whiskers for age')
plt.show()
```


Box plot & whiskers for age

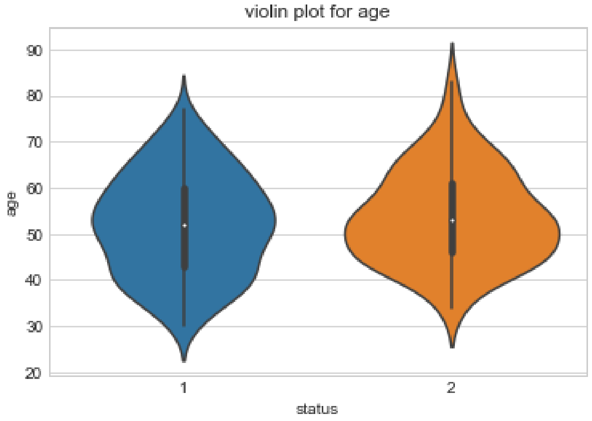


observations: Box plot generally represents in combined form pdf & cdf. 1.) the box plot of status 1 represents. 25th percentile at 42,50th percentile at 53 ,75th percentile at 60. 2.) simllarly for the box plot of status 2. 25th percentile at 47,50th percentile at 57 ,75th percentile at 61.

violin plot

In [234]:

```
# A violin plot combines the boxplot & whiskers  the middle dark line represents compressed boxplo
t
# Denser regions of the data is fatter, and sparser ones thinner
sns.violinplot(x='status',y='age',data=haber)
plt.title('violin plot for age')
plt.show()
```

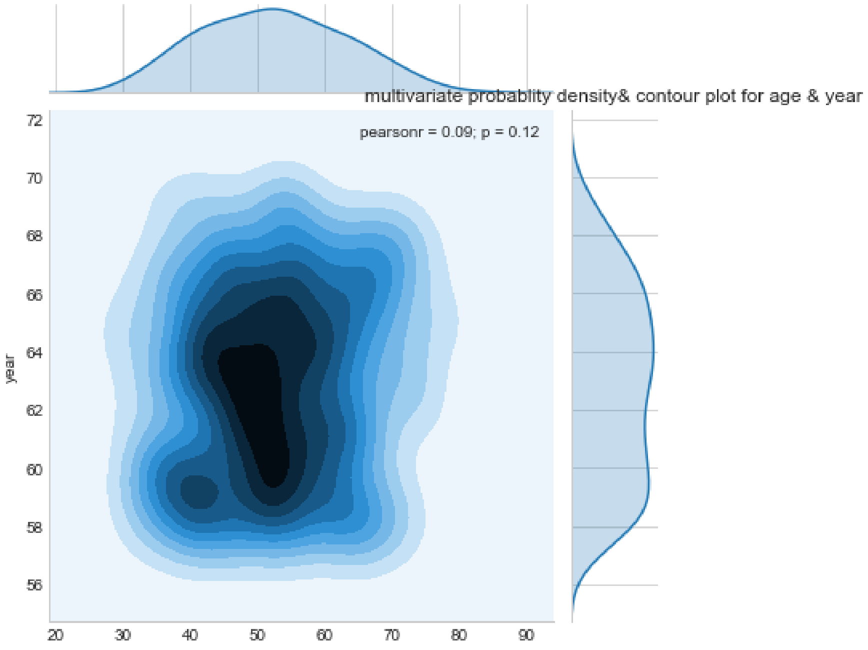


observation: The violin plot inside dark line represents the box plot & whiskers. It generally prints histogram pdf distribution.

Multivariate probability density&contour plot.

In [235]:

```
#2D Density plot, contors-plot
sns.jointplot(x="age", y="year", data=haber, kind="kde");
plt.title('multivariate probablity density& contour plot for age & year')
plt.show()
```



age

observation: 1.) The dark circle inside represents the no of more data points & blue surface area represents conotur. 2.)Distrubtion plots for x and y is also seperated

Result:

1.)The given dataset is imbalanced as it does not contains euqal number of data-points for each class. 2.)The given dataset is not linearly seprable form each class. There are too much overlapping in the data-points and hence it is very diffucult to classify. 3.)we need to have some more complex technique to handle this dataset.