Assignment - Performing EDA on Haberman Cancer Survival dataset

What is Haberman's Cancer Survival dataset?

- The Haberman's Cancer Survival 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.

Objective of Exploratory Data Analysis:

- To classify a patient survival who had undergone surgery for breast cancer.

In [1]:

```
import warnings
warnings.filterwarnings("ignore")
```

In [2]:

```
#code to open the data set
from google.colab import files
uploaded = files.upload()
```

Choose Files No file chosen

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving haberman.csv to haberman.csv
```

In [3]:

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

In [4]:

```
haberman = pd.read_csv("haberman.csv")
```

Observing the Dataset

In [6]:

```
# number of data point & features
print (haberman.shape)
```

```
(306, 4)
```

In [8]:

```
#What are the column names in our dataset?
print (haberman.columns)
```

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

Observation on Dataset:

In the above code we can see their are 4 columns in the dataset, below are the information provided in the respective column.

1. Age - Age of the patient at the time of operation.
2. Year - year when the operation was conducted.
3. Nodes - Number of positive axillary nodes detected. (A axillary lymph nodes are near the breasts, they are often the first location to which breast cancer spreads if it moves beyond the breast tissue. more details on axillary nodes - https://www.medicalnewstoday.com/articles/319713#what-is-the-connection (https://www.medicalnewstoday.com/articles/319713#what-is-the-connection))
4. Status - there are 1 & 2 integers present which describes; 1 = the patient survived 5 years or longer. 2 = the patient died within 5 year

In [9]:

```
#haberman has 2 integers
haberman['status'].unique()
```

Out[9]:

```
array([1, 2])
```

In [10]:

```
#minimum
print(haberman[['age', 'nodes', 'status']].min())
```

```
age       30
nodes      0
status     1
dtype: int64
```

In [11]:

```
#maximum
print(haberman[['age', 'nodes', 'status']].max())
```

```
age       83
nodes     52
status     2
dtype: int64
```

In [12]:

```
#counts of status
haberman.groupby('status').size()
```

Out[12]:

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

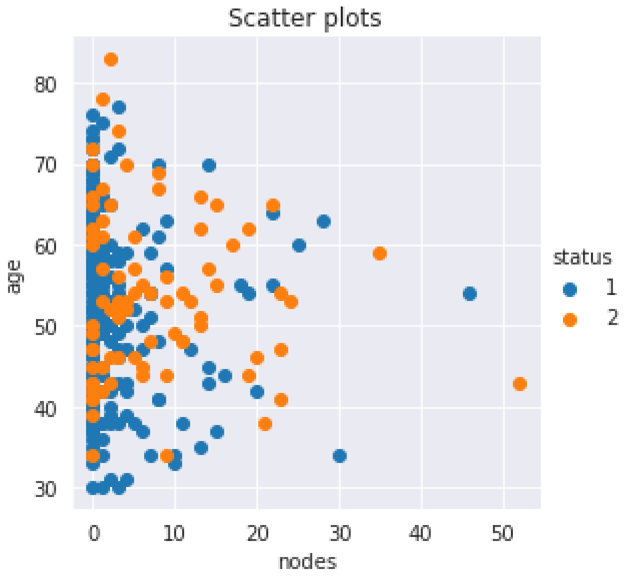
Observation on Dataset:

1. The minimum and maximum age of patient is 30 & 83 respectively.
2. The range of nodes is between 0 to 52.

Scatter Plots

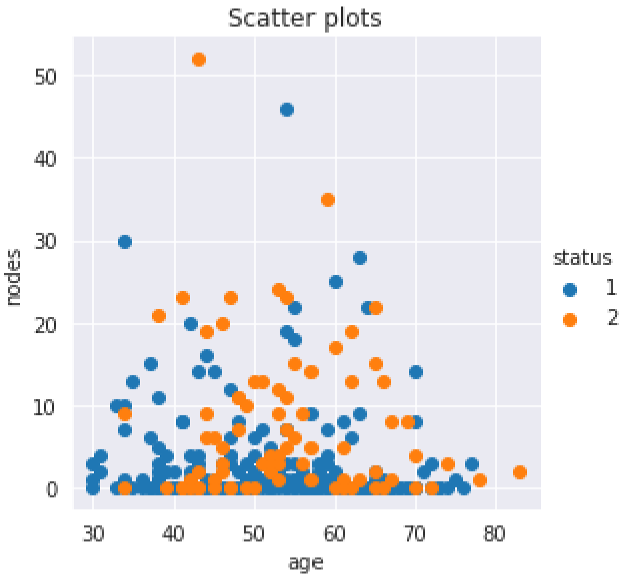
In [32]:

```
#2d scatter plot
sns.set_style("darkgrid")
sns.FacetGrid(haberman, hue="status", height=4) \
.map(plt.scatter, "nodes", "age") \
.add_legend()
plt.title("Scatter plots");
plt.show()
```



In [33]:

```
#2d scatter plot
sns.set_style("darkgrid")
sns.FacetGrid(haberman, hue="status", height=4) \
.map(plt.scatter, "age", "nodes") \
.add_legend()
plt.title("Scatter plots");
plt.show()
```

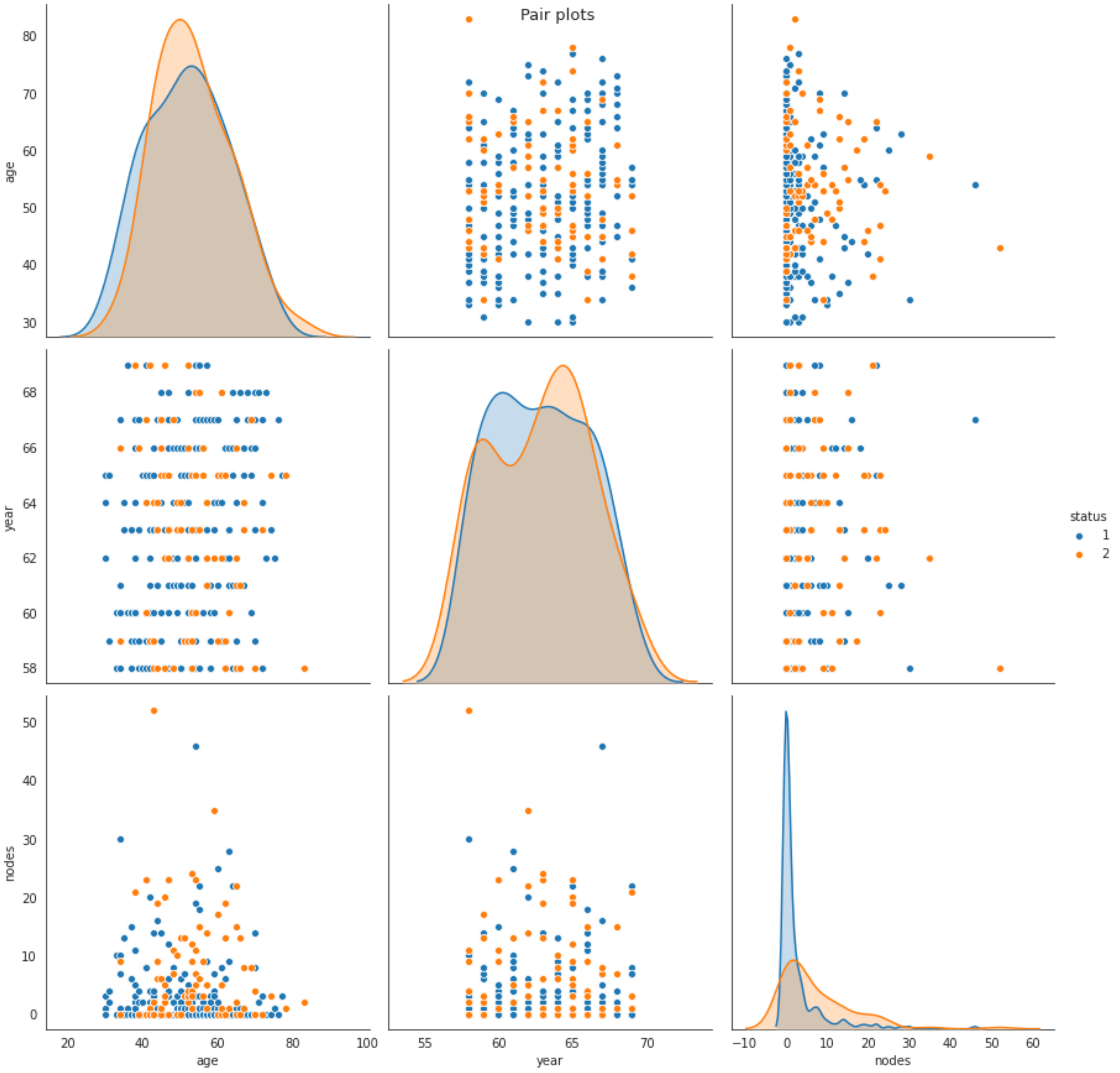


Observations (Scatter plots) :

1. We can identify the survival rate by studying the nodes & age.
2. By using 2D scatter plot we can see patient with nodes between 0 to 2 have chances to survive more than 5 years.

Pair plots

In [15]:
```
plt.close();
sns.set_style("white")
sns.pairplot(haberman, hue="status", height=4)
plt.suptitle("Pair plots",verticalalignment='baseline',horizontalalignment ='c
enter', fontsize=13)
plt.show()
```



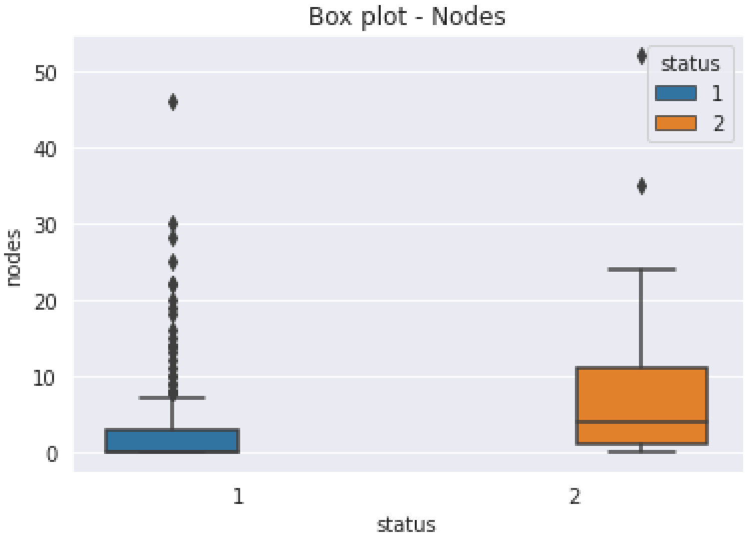
Observation (Pair Plots):

Same observation as Scatter plots.

Box Plot

```
In [16]: status_1 = haberman.loc[haberman["status"] == 1];
         status_2 = haberman.loc[haberman["status"] == 2];
```

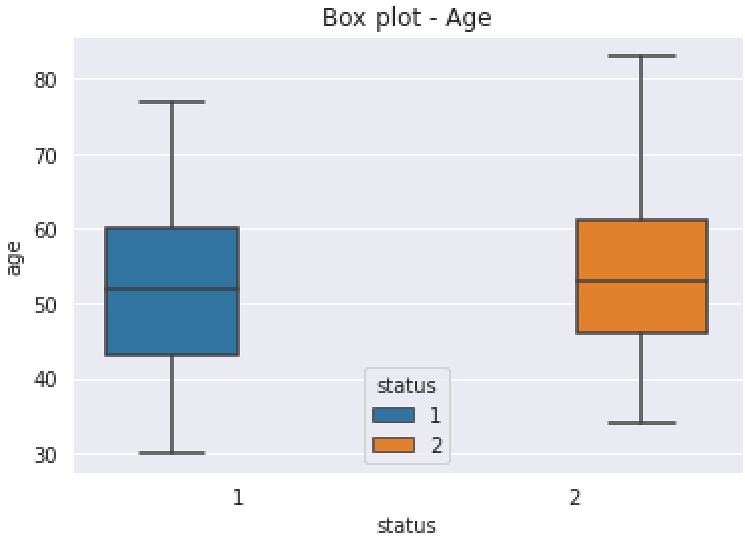
```
In [34]: #box plot & Whiskers
         sns.boxplot(x = "status", y = "nodes", hue = "status", data = haberman)
         plt.title("Box plot - Nodes");
         plt.show()
```



Observation (Box plot)

Chances of surviving more than 5years for less nodes are higher.

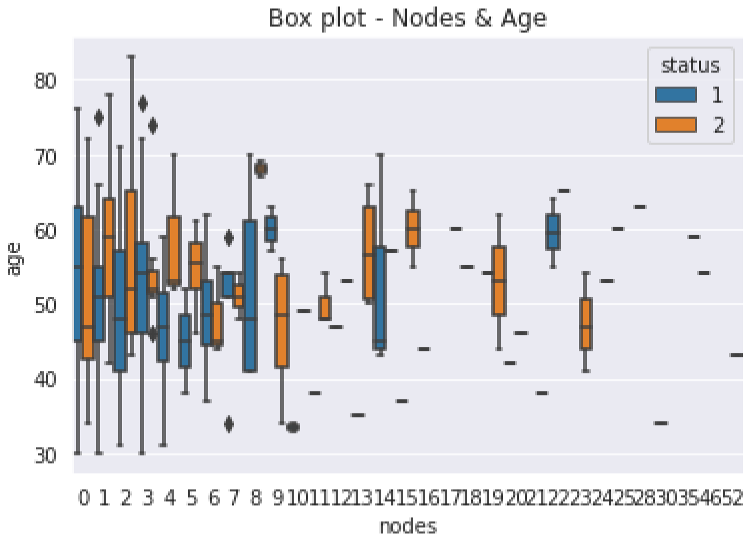
```
In [35]: #box plot & Whiskers
         sns.boxplot(x = "status", y = "age", hue = "status", data = haberman)
         plt.title("Box plot - Age");
         plt.show()
```



Observation (Box Plot)

Median age of both status have almost overlap. Nothing significant information can be derived.

```
In [36]: #box plot & Whiskers
         sns.boxplot(x = "nodes", y = "age", hue = "status", data = haberman, width = 1
         )
         plt.title("Box plot - Nodes & Age");
         plt.show()
```

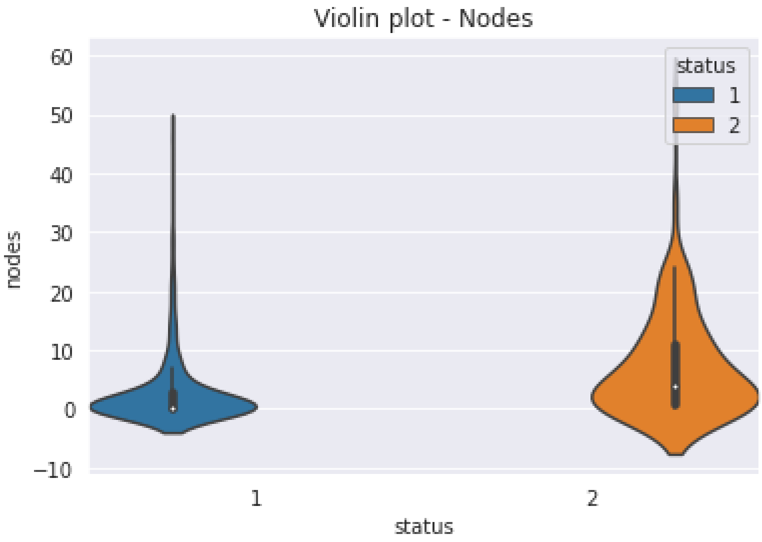


Observations (Box Plot):

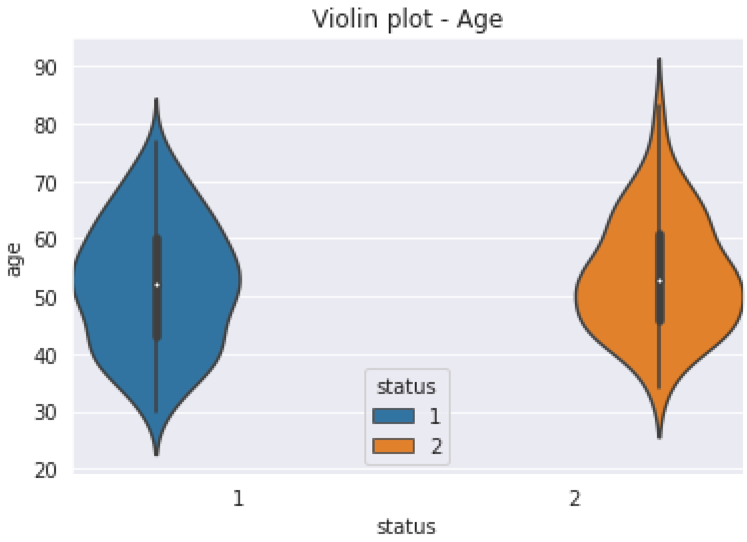
1. I have compared the age and nodes in the above box plots.
2. For every node we can get the range of age group for which good number of data is available.
3. We can see that for less number of nodes, age data range is much larger than higher number of nodes.

Violin Plot

In [37]:
```
#violin plot
sns.violinplot(x="status", y="nodes", hue="status", data=haberman, width=1 )
plt.title("Violin plot - Nodes");
plt.show()
```



In [38]:
```
#violin plot
sns.violinplot(x="status", y="age", hue= "status", data=haberman, width=1)
plt.title("Violin plot - Age");
plt.show()
```

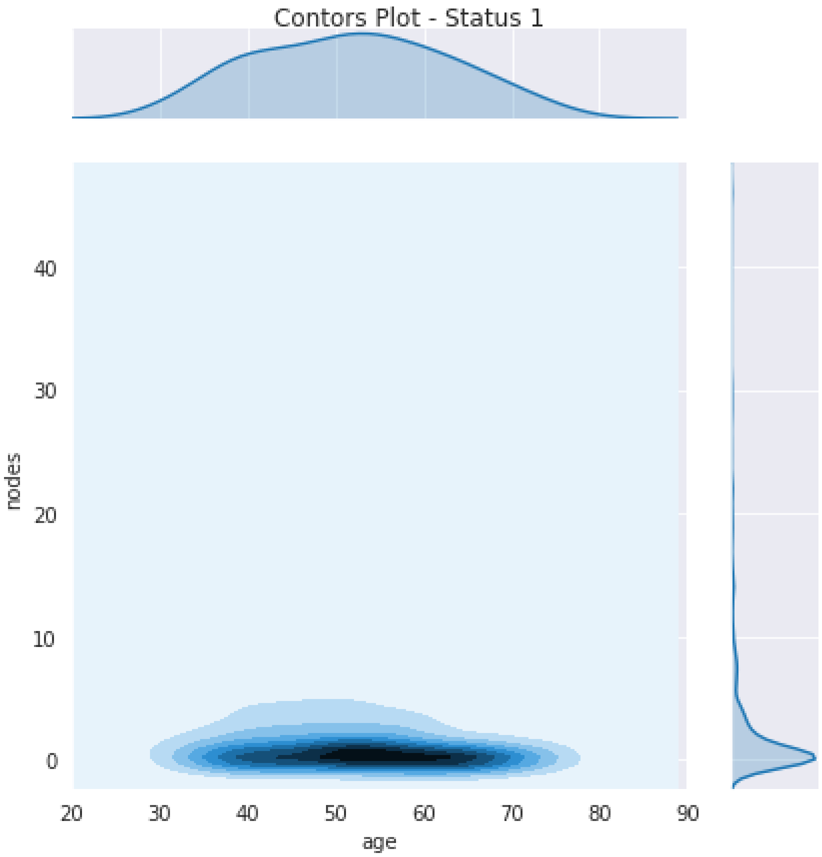


Observations (violin plot):

1. Through violin plots we can observe the survival status is more if the nodes are less.
2. We can't get proper observation by only age data, we will require nodes data.

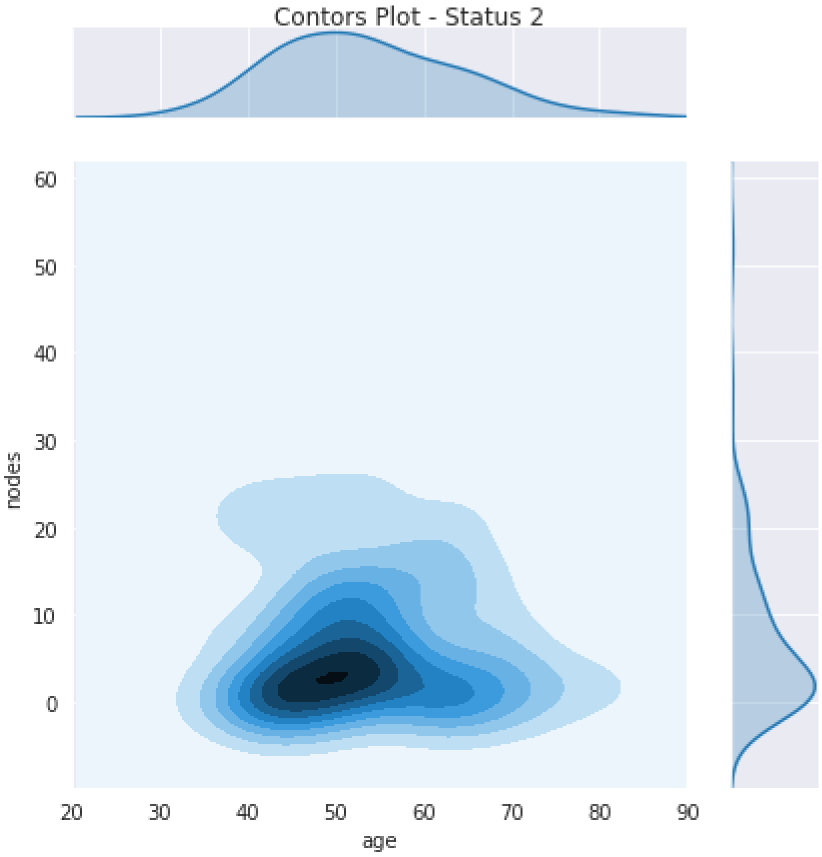
Contors Plot

In [63]:
```
#contors-plot
sns.jointplot(x="age", y="nodes", data=status_1, kind="kde",space=0.5,xlim= (20, 90) )
plt.suptitle("Contors Plot - Status 1", verticalalignment='baseline')
plt.show();
```



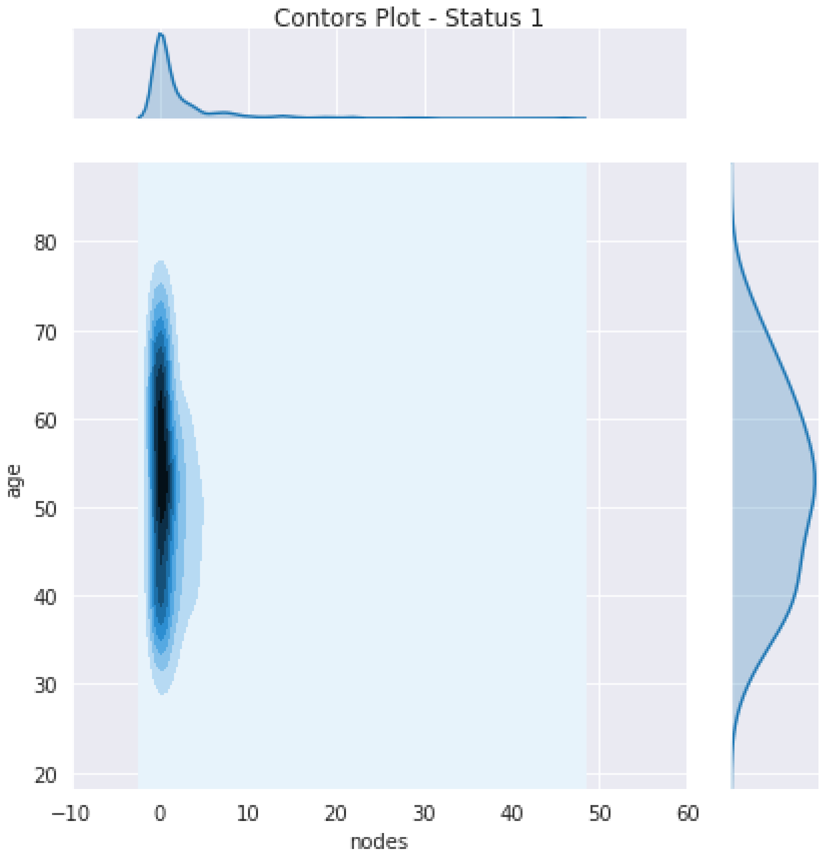
In [64]:

```
#contors-plot
sns.jointplot(x="age", y="nodes", data=status_2, kind="kde",space=0.5,xlim= (20, 90) );
plt.suptitle("Contors Plot - Status 2", verticalalignment='baseline')
plt.show();
```

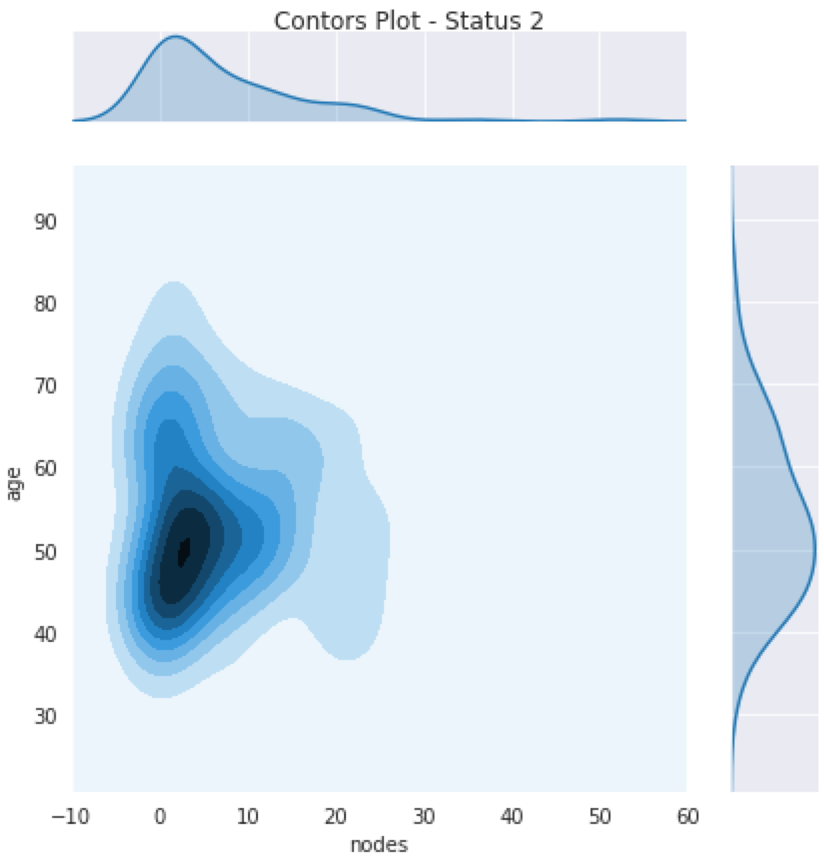


In [65]:

```
#contors-plot
sns.jointplot(x="nodes", y="age", data=status_1, kind="kde",space=0.5,xlim= (-
10, 60) )
plt.suptitle("Contors Plot - Status 1", verticalalignment='baseline');
plt.show();
```



```
In [66]: #contors-plot
         sns.jointplot(x="nodes", y="age", data=status_2, kind="kde",space=0.5,xlim= (-
         10, 60) )
         plt.suptitle("Contors Plot - Status 2", verticalalignment='baseline');
         plt.show();
```



Observations (contors-plot):

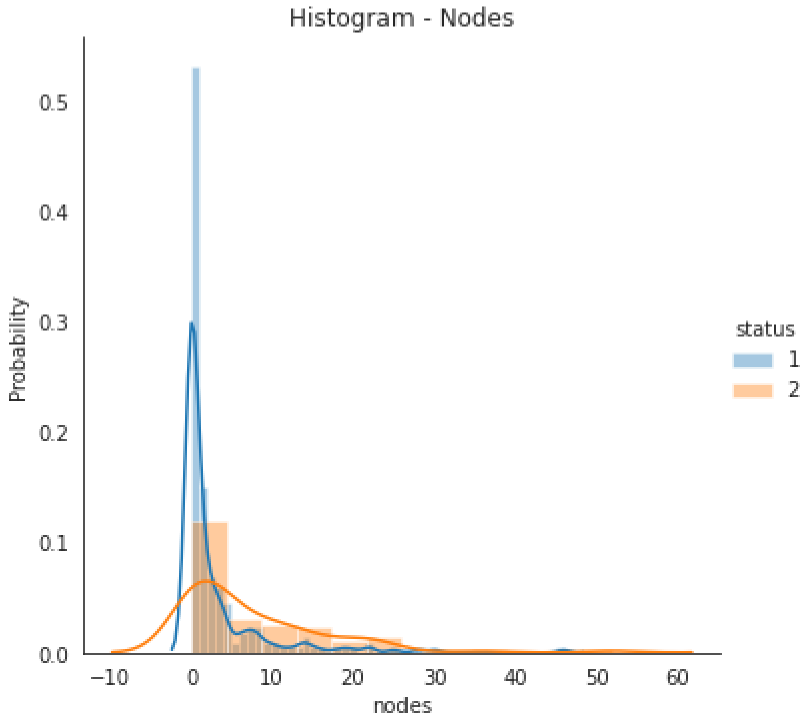
By studying the relation between nodes and age we can see the below,

1. Patient with nodes between 0 to 2 have survived more than 5years.
2. Patient of age group between 45 to 65 have survived more than 5 years with nodes between 0 to 2.
   3.Patient with more than 5 nodes have mostly survived less than 5 years.

Histogram

In [26]:

```
#histogram
sns.FacetGrid(haberman, hue="status", size=5) \
   .map(sns.distplot, "nodes") \
   .add_legend()
plt.title("Histogram - Nodes")
plt.ylabel("Probability")
plt.show();
```

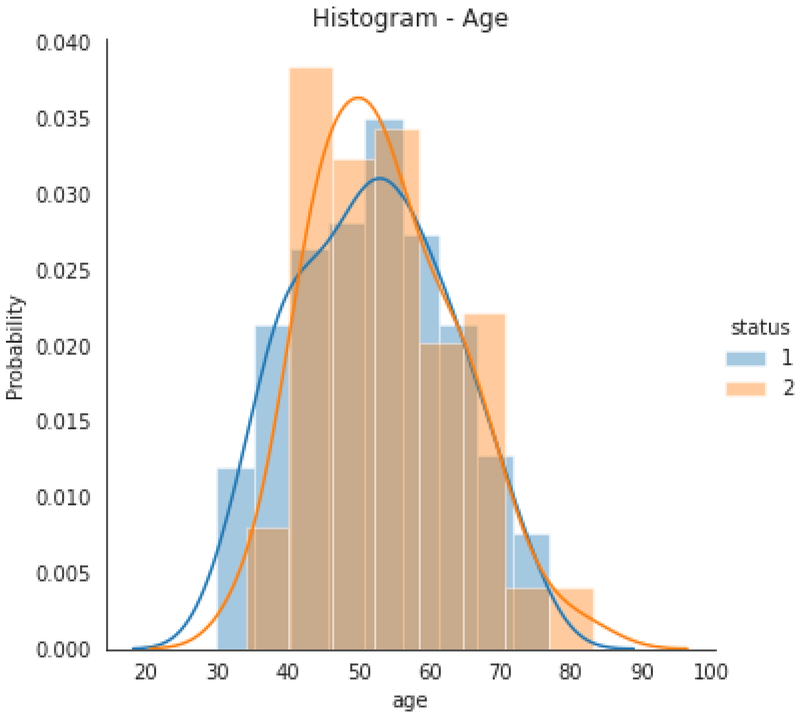


Observation(Histogram - nodes):

As number of nodes increases, probability of surviving more than 5 years decreases.

In [27]:

```
#histogram
sns.FacetGrid(haberman, hue="status", size=5) \
   .map(sns.distplot, "age") \
   .add_legend()
plt.title("Histogram - Age")
plt.ylabel("Probability")
plt.show();
```



Observation(Histogram - age):

There's overlapping of bars, hence we can't analyse the survival status by just using age data.

Pdf & Cdf

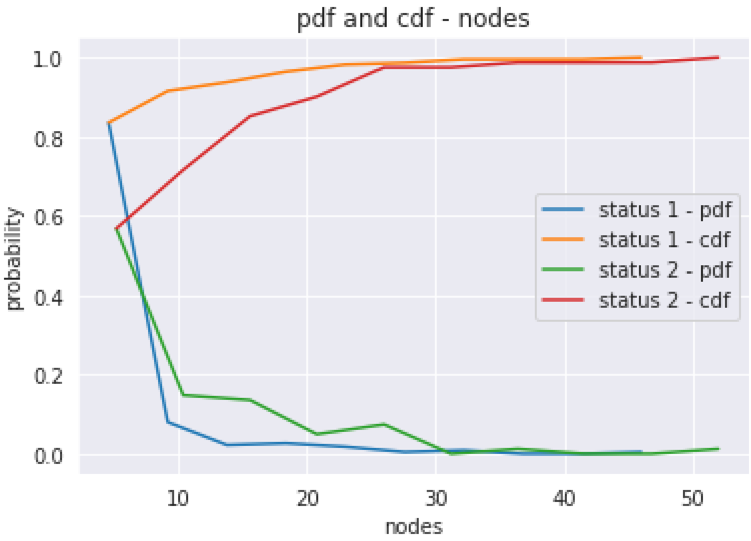
In [62]:

```python
#pdf cdf
counts, bin_edges = np.histogram(status_1['nodes'], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)

counts, bin_edges = np.histogram(status_2['nodes'], bins=10,
                                 density = True)
pdf = counts/(sum(counts))
print(pdf);
print(bin_edges)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)

plt.title("pdf and cdf - nodes")
plt.xlabel("nodes")
plt.ylabel("probability")
legend_status = ["status 1 - pdf", "status 1 - cdf", "status 2 - pdf", "status
2 - cdf"]
plt.legend(legend_status)
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. ]
```



Observation (pdf & cdf):

As number of nodes increases, probability of surviving more than 5 years decreases.

Statistics

In [30]:

```
#mean std dev
print("Means of nodes:")
print(np.mean(status_1["nodes"]))
print(np.mean(status_2["nodes"]))
print("\nStd-dev of nodes:");
print(np.std(status_1["nodes"]))
print(np.std(status_2["nodes"]))
print("\nMeans of age:")
print(np.mean(status_1["age"]))
print(np.mean(status_2["age"]))
print("\nStd-dev of age:");
print(np.std(status_1["age"]))
print(np.std(status_2["age"]))
```

```
Means of nodes:
2.791111111111111113
7.45679012345679

Std-dev of nodes:
5.857258449412131
9.128776076761632

Means of age:
52.01777777777778
53.67901234567901

Std-dev of age:
10.98765547510051
10.10418219303131
```

In [31]:

```
#median quantiles
print("\nMedians of nodes:")
print(np.median(status_1["nodes"]))
print(np.median(status_2["nodes"]))
print("\nMedians of age:")
print(np.median(status_1["age"]))
print(np.median(status_2["age"]))
print("\nQuantiles of nodes:")
print(np.percentile(status_1["nodes"],np.arange(0, 100, 25)))
print(np.percentile(status_2["nodes"],np.arange(0, 100, 25)))
print("\nQuantiles of age:")
print(np.percentile(status_1["age"],np.arange(0, 100, 25)))
print(np.percentile(status_2["age"],np.arange(0, 100, 25)))
```

```
Medians of nodes:
0.0
4.0

Medians of age:
52.0
53.0

Quantiles of nodes:
[0. 0. 0. 3.]
[ 0.  1.  4. 11.]

Quantiles of age:
[30. 43. 52. 60.]
[34. 46. 53. 61.]
```

Observations (Median & Nodes):

1. By analysing the medians & quantiles we can see the patient with nodes less than 3 have chance to survive more than 5 years.
2. Information about nodes is necessary to get clear understanding of survival chances as compared to age.

Conclusion:

1. Through various analysis we can conclude that the nodes data is necessary in identifying survival status.
2. By using box plots we can relate the age range with number of nodes in a better way. This can help in identifying the survival status of age group against the nodes number.