In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:
```
df = pd.read_csv('haberman.csv')
df.head()
```

Out[4]:

	age	year	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 [7]:
```
df.shape
```

Out[7]:
```
(306, 4)
```

- This dataset has 306 datapoints.
- This Dataset has 4 attribute columns.

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306 entries, 0 to 305
Data columns (total 4 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   age     306 non-null    int64
 1   year    306 non-null    int64
 2   nodes   306 non-null    int64
 3   status  306 non-null    int64
dtypes: int64(4)
memory usage: 9.7 KB
```

- We have 4 attributes all of integer type
- There are no missing values
- In status column, 1 represents Survived at least 5 years , 2 represents survived less than 5 years

In [8]:
```
df.describe()
```

Out[8]:

	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

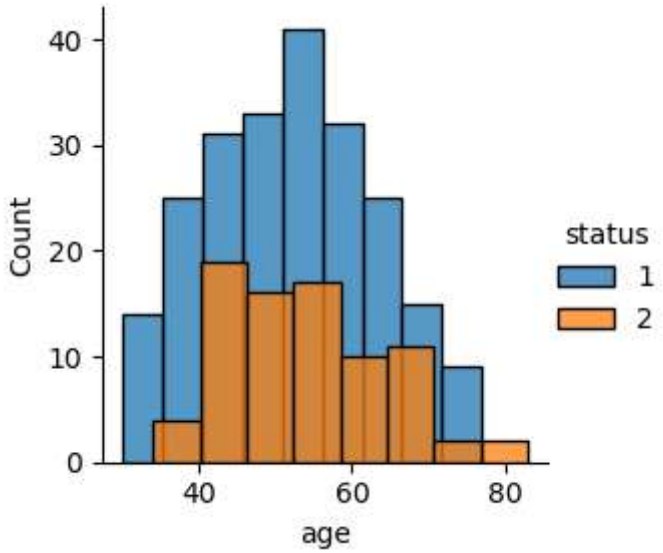
- Year column has no meaning full data, Since that is year of operation
- Avg age of patients is 52.5
- About 75% of patients has at most 4 nodes
- About 25% of patients has no nodes

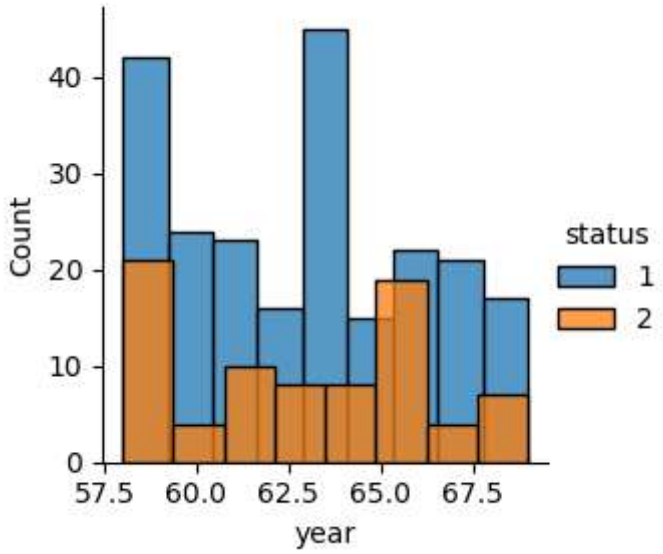
In [17]:
```
(df['status'] == 1).value_counts()
```

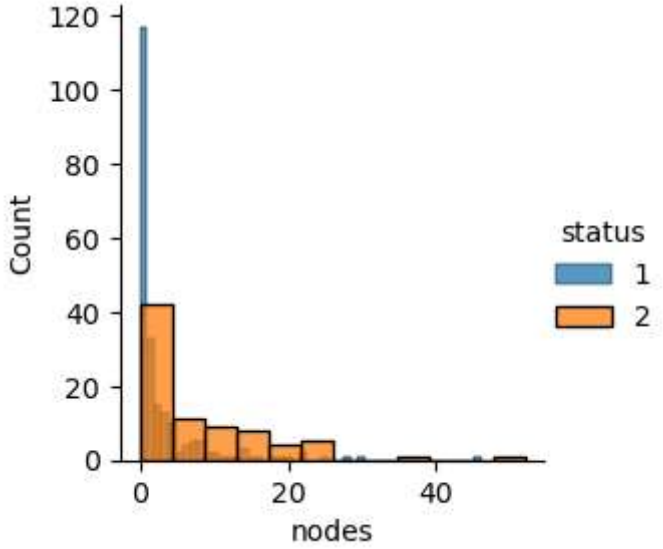
Out[17]:
```
True     225
False     81
Name: status, dtype: int64
```

- This is moderately Skewed Dataset with 73.5% Survived for at least 5 years datapoints.

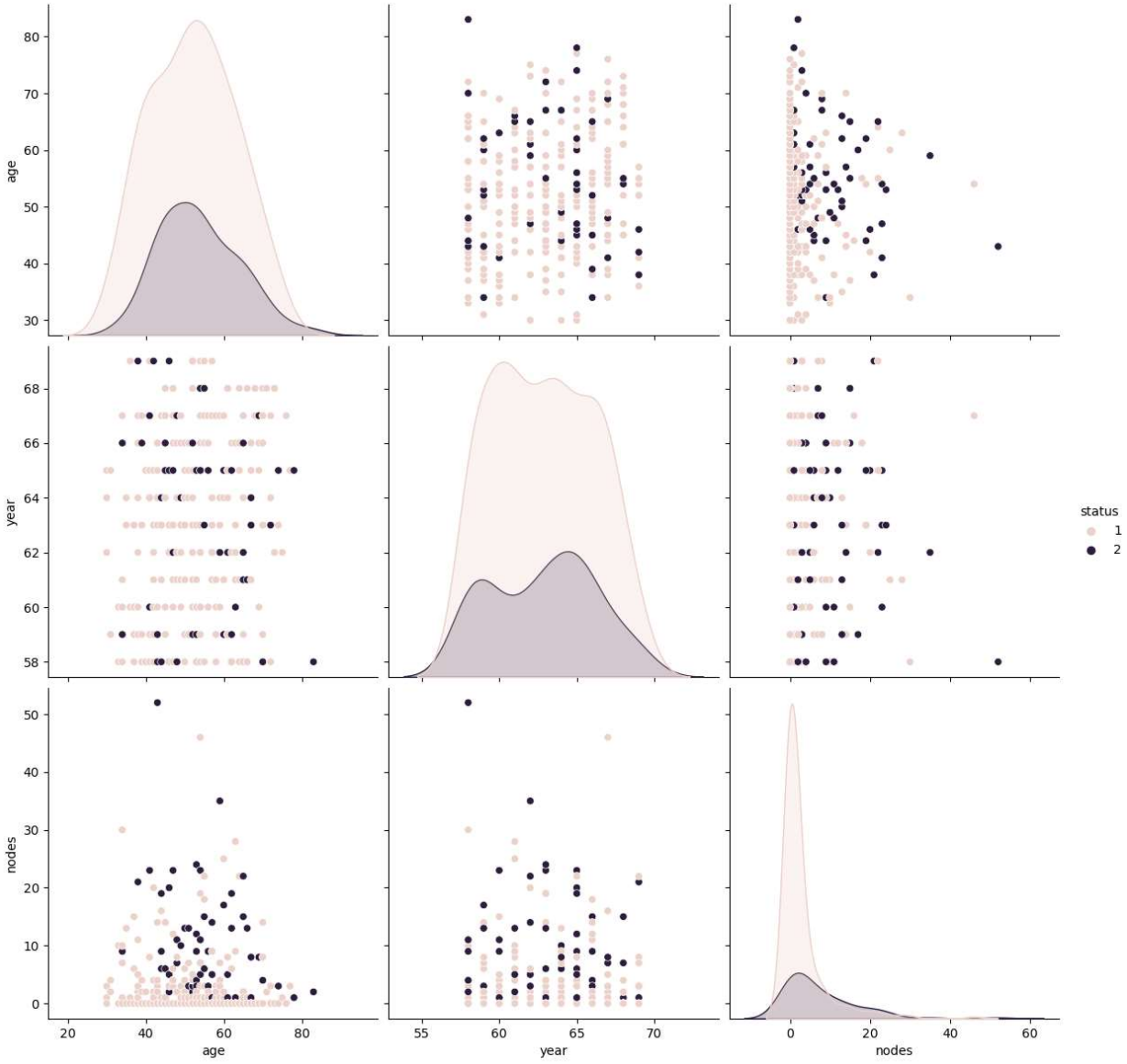
In [25]:
```
for feature in list(df.columns)[:-1]:
    graph = sns.FacetGrid(df,hue = 'status',height= 3)
    graph.map(sns.histplot,feature).add_legend()
    plt.show()
```







In [27]:
```
sns.pairplot(df, hue='status', height=4)
plt.show()
```



In [39]:
```
patients40 = df[df['age'] < 40]['status'].value_counts()
print(patients40)
```

```
1    36
2     4
Name: status, dtype: int64
```

- Irrespective of number of nodes, Patients survived for at least 5 years in 90 % of time (Sample data is small with age < 40 so we can not say with confidance) .