```
import warnings
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
warnings.filterwarnings("ignore")
%matplotlib inline
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
     Drive already mounted at /content/drive; to attempt to forcibly remount, call
```

```
cd /content/drive/MyDrive/AAIC/
```

```
     /content/drive/MyDrive/AAIC
```

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

```
data.head(5)
```

	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

```
data.shape
```

```
     (306, 4)
```

Observation

1. This data set has 306 data points
2. Data set has four features

```
print(data.columns)
```

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

Observation There are two unique survival status

1. Staus is in integer format which is not clear so we will convert them
2. We will replace survival_staus=1 to yes and survival_status=2 to no

```
data["survival_status"].replace({1:"yes",2:"no"},inplace=True)
```

```
data.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   survival_status  306 non-null    object
dtypes: int64(3), object(1)
memory usage: 9.7+ KB
```

Observation

1. No cloumns has missing datapoints
2. three of the data points are of type integer
3. Survival_status which was of type integer is now of type obect

```
data.describe()
```

	age	year	nodes
count	306.000000	306.000000	306.000000
mean	52.457516	62.852941	4.026144
std	10.803452	3.249405	7.189654
min	30.000000	58.000000	0.000000

- Maximum no of nodes =52
- Age vary between 30 to 83

```
data["survival_status"].isnull().value_counts()
```

```
False    306
Name: survival_status, dtype: int64
```

Observation

1. We don't have any missing values

```
data["survival_status"].value_counts()
```

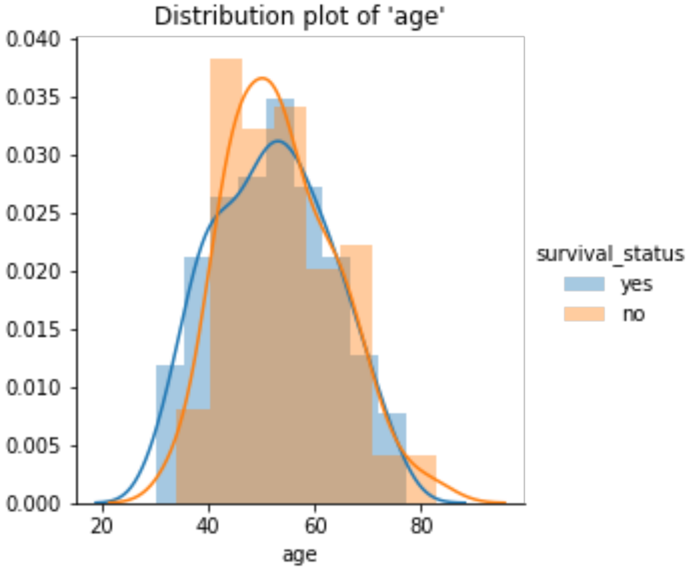
```
yes    225
no      81
Name: survival_status, dtype: int64
```

ovservation

- out of 306 patients 225 patients survived and 81 patients didn't survive
- also we can say dataset is imbalanced

```
sns.FacetGrid(data,hue="survival_status",height=4).map(sns.distplot,"age").add_lege
```

```
<seaborn.axisgrid.FacetGrid at 0x7f2e50ad1d90>
```



```
print(sur.shape)
total_not_sur=data[data["survival_status"]=="no"]
print(total_not_sur.shape)
not_sur=total_not_sur[(total_not_sur["age"]>=40) & (total_not_sur["age"]<=57)]
print(not_sur.shape)
percentage_sur_40_57=(sur.shape[0]/total_sur.shape[0])*100
percentage_not_sur_40_57=(not_sur.shape[0]/total_not_sur.shape[0])*100
print("percentage of beople survived between age 40 and 57 is {}".format(percentage
print("percentage of people not survived between age 40 and 57 is {}".format(percen
```

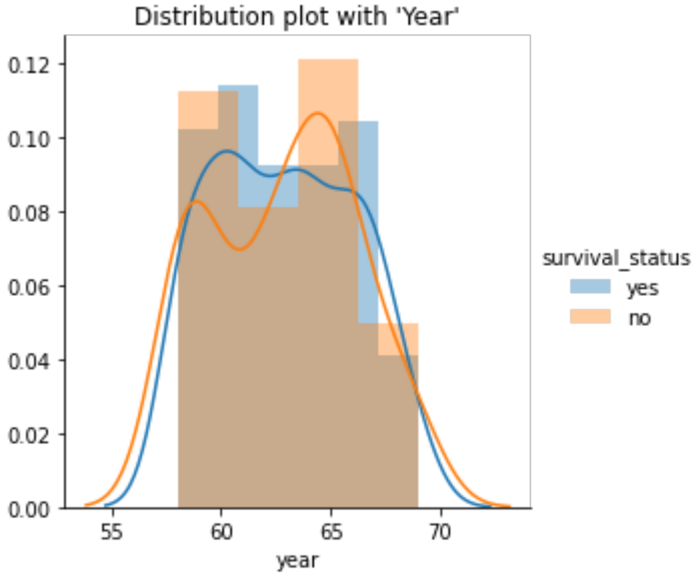
```
(225, 4)
(116, 4)
(81, 4)
(52, 4)
percentage of beople survived between age 40 and 57 is 51.55555555555556
percentage of people not survived between age 40 and 57 is 64.19753086419753
```

Observations

- from above figure we saw that between age group 40 and 57 less people could survive
- Once we calculated that we could see 51.6 % of people survived and 64.2 % people died who all belong to age group 40 and 57

```
sns.FacetGrid(data,hue="survival_status",height=4).map(sns.distplot,"year").add_leg
```

```
<seaborn.axisgrid.FacetGrid at 0x7f2e50a6fc90>
```



```
print(t2.shape)
print(t3.shape)
print(t4.shape)
patients_not_survived_in_1965=(t2.shape[0]/total_not_sur.shape[0])*100
patients_survived_in_1965=(t4.shape[0]/total_sur.shape[0])*100
print("{}% patients could survive in year 1965 ".format(int(patients_survived_in_19
print("{}% patients could not survive in year 1965".format(int(patients_not_survive
```

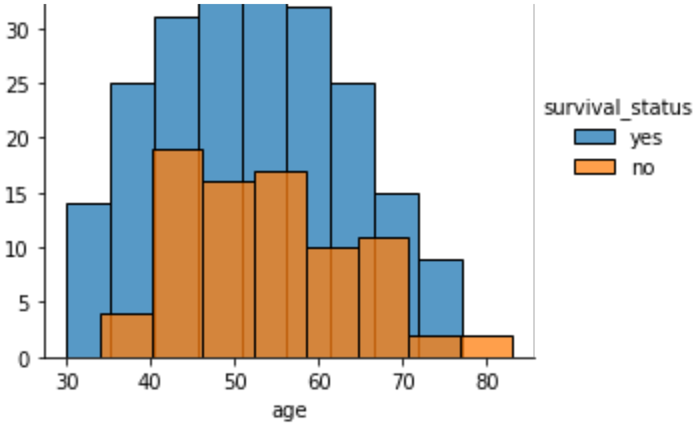
```
    (8, 4)
    (13, 4)
    (22, 4)
    (15, 4)
    6% patients could survive in year 1965
    16% patients could not survive in year 1965
```

Observation

- 16% patients didn't survive in 1965
- only 6% of patients could survive in year 1965
- there may be some medical challenges which led more death in year 1965

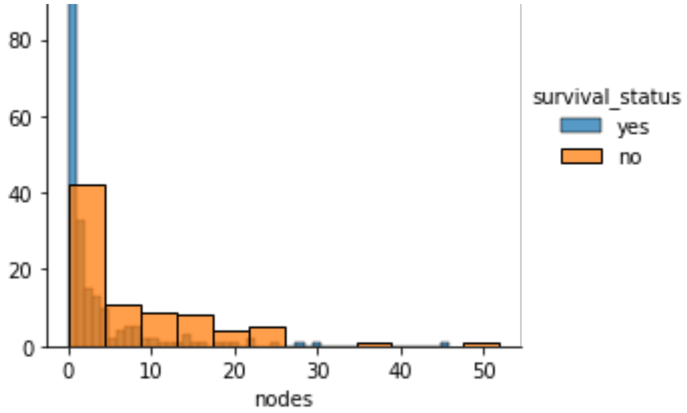
```
sns.FacetGrid(data,hue="survival_status",height=4).map(sns.distplot,"nodes").add_le
```

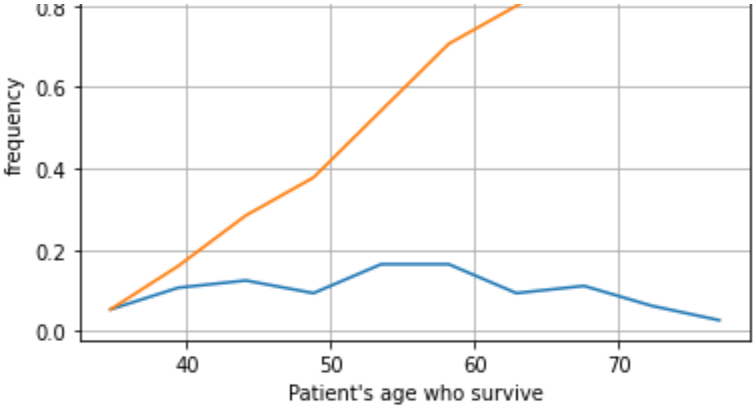
```
    <seaborn.axisgrid.FacetGrid at 0x7f2e5109abd0>
```



My Observation

- After age 75 no patients could survive
- Patients of age between 30 to 33 could survive there is no death can be found from histogram plot in this age group


80
60
40
20
0
0
10
20
30
40
50
nodes
survival_status
yes
no


0.8
0.6
0.4
0.2
0.0
frequency
40
50
60
70
Patient's age who survive