Exploratory Data Analysis (EDA)

Importing libraries and loading data

In [1]:

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

In [2]:

```
df=pd.read_csv("D:/Downloads/archive (21)/haberman.csv",header=0)
df.columns=['patient_age', 'operation_year', 'positive_axillary_nodes', 'sur
df
```

Out[2]:

	patient_age	operation_year	positive_axillary_nodes	survival_status
0	30	62	3	1
1	30	65	0	1
2	31	59	2	1
3	31	65	4	1
4	33	58	10	1
...	...	...	...	...
300	75	62	1	1
301	76	67	0	1
302	77	65	3	1
303	78	65	1	2
304	83	58	2	2

305 rows × 4 columns

Data understanding

In [3]:

```
df.shape
```

Out[3]:

```
(305, 4)
```

In [4]:

```
df['survival_status'].value_counts()
```

Out[4]:

```
survival_status
1    224
2     81
Name: count, dtype: int64
```

```
The dataset is imbalanced, out of total of 305 patients no. of survived is
3 times the patients who died within 5 years
```

In [5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 305 entries, 0 to 304
Data columns (total 4 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   patient_age              305 non-null    int64
 1   operation_year           305 non-null    int64
 2   positive_axillary_nodes  305 non-null    int64
 3   survival_status          305 non-null    int64
dtypes: int64(4)
memory usage: 9.7 KB
```

In [6]:

```
#output shows all integer non null values
```

In [7]:

```
df['survival_status'] = df['survival_status'].map({1:"yes", 2:"no"})
df
```

Out[7]:

	patient_age	operation_year	positive_axillary_nodes	survival_status
0	30	62	3	yes
1	30	65	0	yes
2	31	59	2	yes
3	31	65	4	yes
4	33	58	10	yes
...	...	...	...	...
300	75	62	1	yes
301	76	67	0	yes
302	77	65	3	yes
303	78	65	1	no
304	83	58	2	no

305 rows × 4 columns

In [8]:

```
df.describe()
```

Out[8]:

	patient_age	operation_year	positive_axillary_nodes
count	305.000000	305.000000	305.000000
mean	52.531148	62.849180	4.036066
std	10.744024	3.254078	7.199370
min	30.000000	58.000000	0.000000
25%	44.000000	60.000000	0.000000
50%	52.000000	63.000000	1.000000
75%	61.000000	66.000000	4.000000
max	83.000000	69.000000	52.000000

```
patients got operated at age of 63
average number of positive axillary nodes detected =4
50th percentile, the median of positive axillary nodes is 1.
75th percentile, 75% of the patients have less than 4 nodes detected.
"there is a significant difference between the mean and the median
values(50%). This is because there are some outliers in our data and the
mean is influenced by the presence of outliers.It indicate potential
outliers, it's not conclusive proof.Or (mean>median) skewed"
(positive_axillary_nodes mean =4 and median=1 difference is high)
```

Class-wise statistical analysis

In [9]:

```
survival_yes=df[df['survival_status']=='yes']
survival_yes.describe()
```

Out[9]:

	patient_age	operation_year	positive_axillary_nodes
count	224.000000	224.000000	224.000000
mean	52.116071	62.857143	2.799107
std	10.937446	3.229231	5.882237
min	30.000000	58.000000	0.000000
25%	43.000000	60.000000	0.000000
50%	52.000000	63.000000	0.000000
75%	60.000000	66.000000	3.000000
max	77.000000	69.000000	46.000000

In [10]:
```
survival_no=df[df['survival_status']=='no']
survival_no.describe()
```

Out[10]:

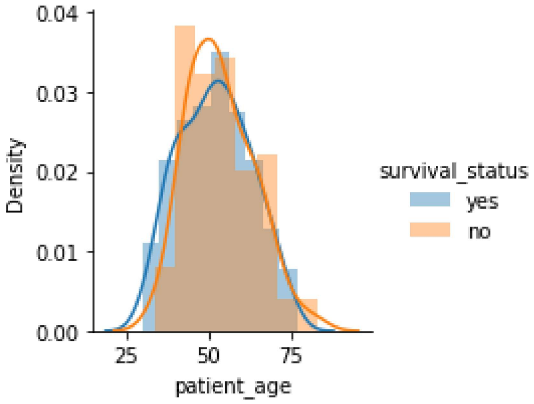
	patient_age	operation_year	positive_axillary_nodes
count	81.000000	81.000000	81.000000
mean	53.679012	62.827160	7.456790
std	10.167137	3.342118	9.185654
min	34.000000	58.000000	0.000000
25%	46.000000	59.000000	1.000000
50%	53.000000	63.000000	4.000000
75%	61.000000	65.000000	11.000000
max	83.000000	69.000000	52.000000

```
1.patient is operated on is nearly the same in both cases

2.patient who died within 5 years on average about 4 to 5 positive
axiliary nodes more than patients who lived
```

3. Uni-variate data analysis

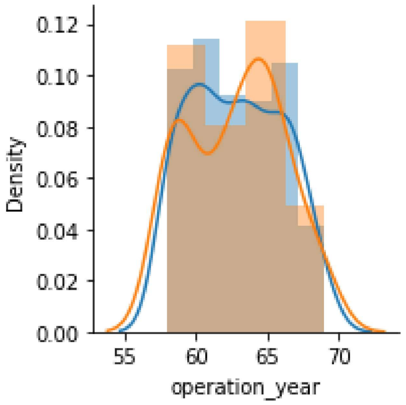
In [11]:
```
with warnings.catch_warnings():
    warnings.simplefilter("ignore")
    sns.FacetGrid(df,hue="survival_status").map(sns.distplot,"patient_age").
    plt.figure(figsize=(15, 8))
    plt.show()
```



```
<Figure size 1080x576 with 0 Axes>
```

```
Among all the age groups, the patients belonging to 40-60 years of age are
the highest
```

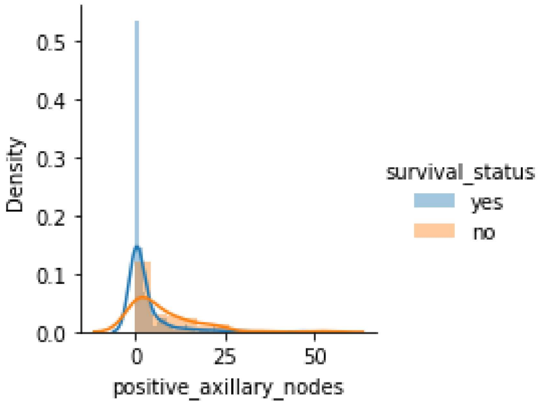
In [12]:
```
with warnings.catch_warnings():
    warnings.simplefilter("ignore")
    sns.FacetGrid(df, hue = "survival_status").map(sns.distplot, "operation_
    plt.show()
```



Huge overlap between the class labels suggesting that one cannot make any distinctive conclusion regarding the survival status based solely on the operation year and patient's age.

Number of positive axillary nods

In [13]:
```
with warnings.catch_warnings():
    warnings.simplefilter("ignore")
    g = sns.FacetGrid(df, hue="survival_status")
    g.map(sns.distplot, "positive_axillary_nodes", kde=True)
    g.add_legend()
    plt.figure(figsize=(12,6))
    plt.show()
```



```
<Figure size 864x432 with 0 Axes>
```

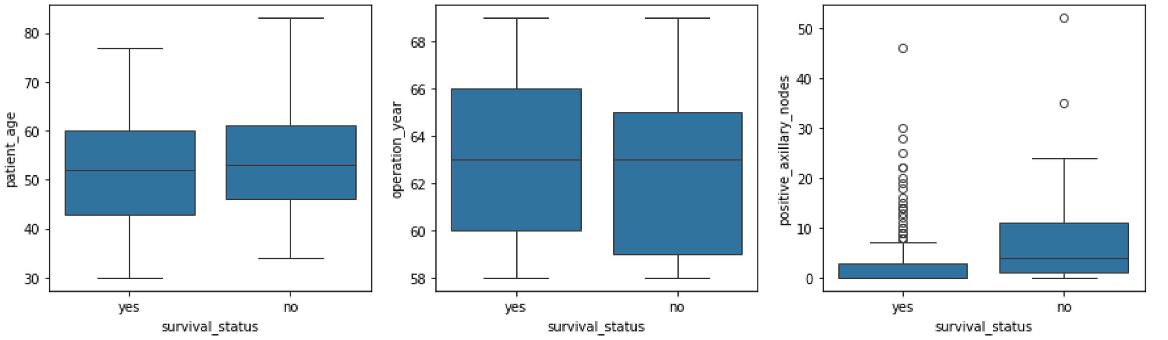
Patients having 4 or fewer axillary nodes — A very good majority of these patients have survived 5 years

Box plot

```
The box plot, commonly referred to as a box and whisker plot, serves as a
visual representation that summarizes exploratory data analysis Python
using five key metrics – the minimum, lower quartile (25th percentile),
median (50th percentile), upper quartile (75th percentile), and maximum
data values.
```

In [14]:

```
plt.figure(figsize = (15, 4))
plt.subplot(1,3,1)
sns.boxplot(x = 'survival_status', y = 'patient_age', data = df)
plt.subplot(1,3,2)
sns.boxplot(x = 'survival_status', y = 'operation_year', data = df)
plt.subplot(1,3,3)
sns.boxplot(x = 'survival_status', y = 'positive_axillary_nodes', data = df)
plt.show()
```

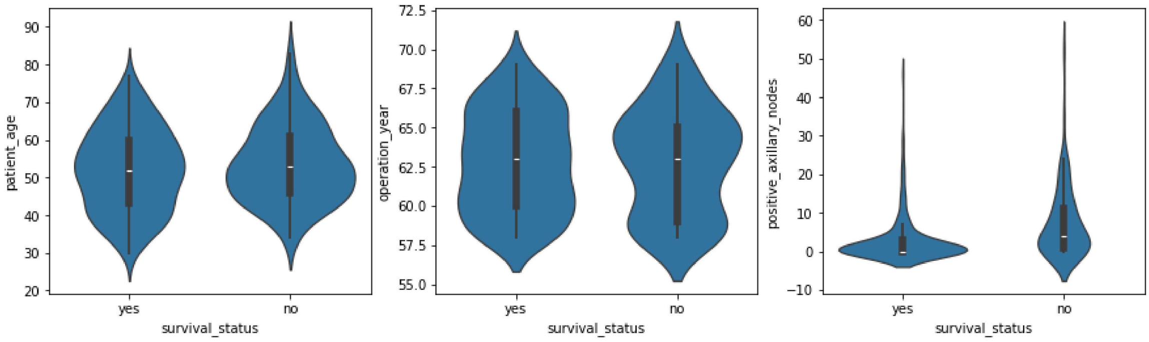


```
patient age and the operation year plots show similar statistics
The isolated points seen in the box plot of positive axillary nodes are
the outliers in the data. Such a high number of outliers is kind of
expected in medical datasets.
```

Violin plot

```
A violin plot displays the same information as the box and whisker plot;
additionally, it also shows the density-smoothed plot of the underlying
distribution.
```

In [15]:
```
plt.figure(figsize = (15, 4))
plt.subplot(1,3,1)
sns.violinplot(x = 'survival_status', y = 'patient_age', data = df)
plt.subplot(1,3,2)
sns.violinplot(x = 'survival_status', y = 'operation_year', data = df)
plt.subplot(1,3,3)
sns.violinplot(x = 'survival_status', y = 'positive_axillary_nodes', data =
plt.show()
```

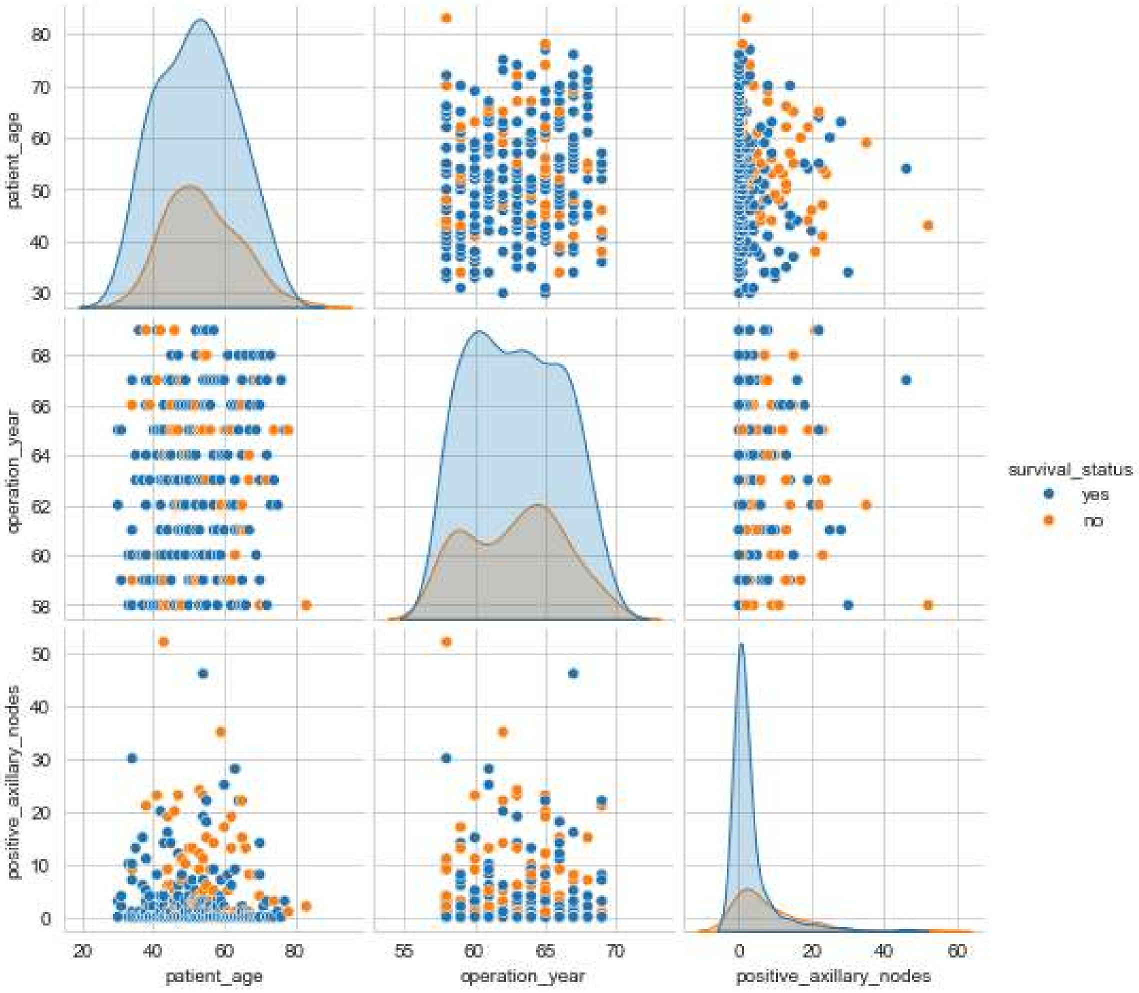


violin plot for positive axillary nodes, it becomes apparent that the distribution is highly skewed for the ‘yes’ class label and moderately skewed for the ‘no’ label.

Bi-variate data analysis

In [16]:

```
#pair plot
sns.set_style('whitegrid')
sns.pairplot(df,hue='survival_status')
plt.show()
```



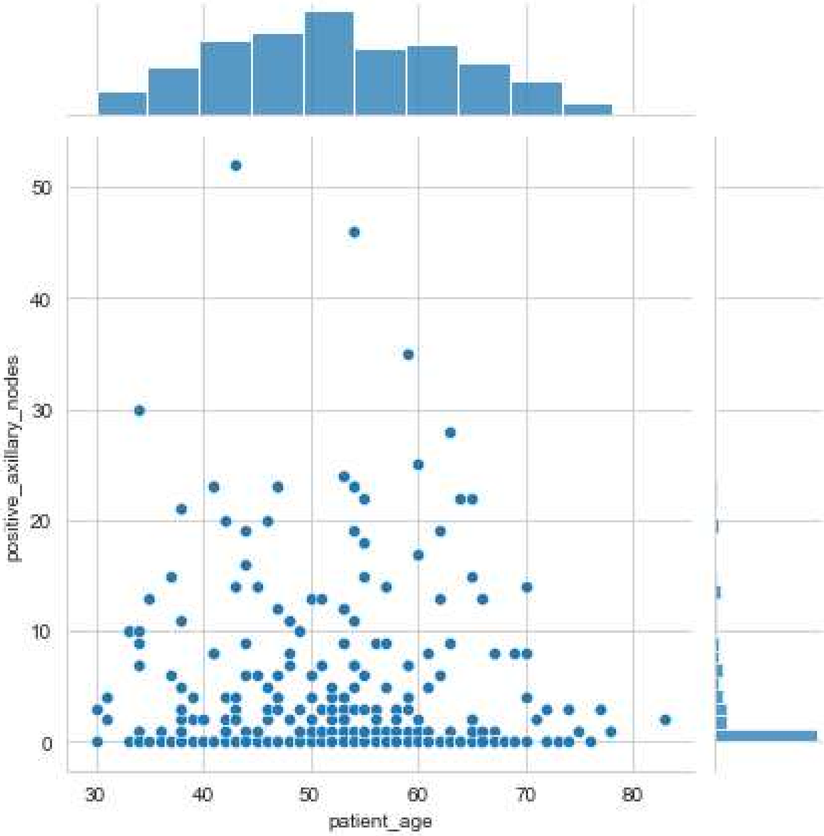
As we can observe in the above pair plot, there is a high overlap between any two features and hence no clear distinction can be made between the class labels

Joint plot

While the Pair plot provides a visual insight into all possible correlations, the Joint plot provides bivariate plots with univariate marginal distributions.

In [17]:

```
sns.jointplot(x="patient_age",y="positive_axillary_nodes",data=df)
plt.show()
```

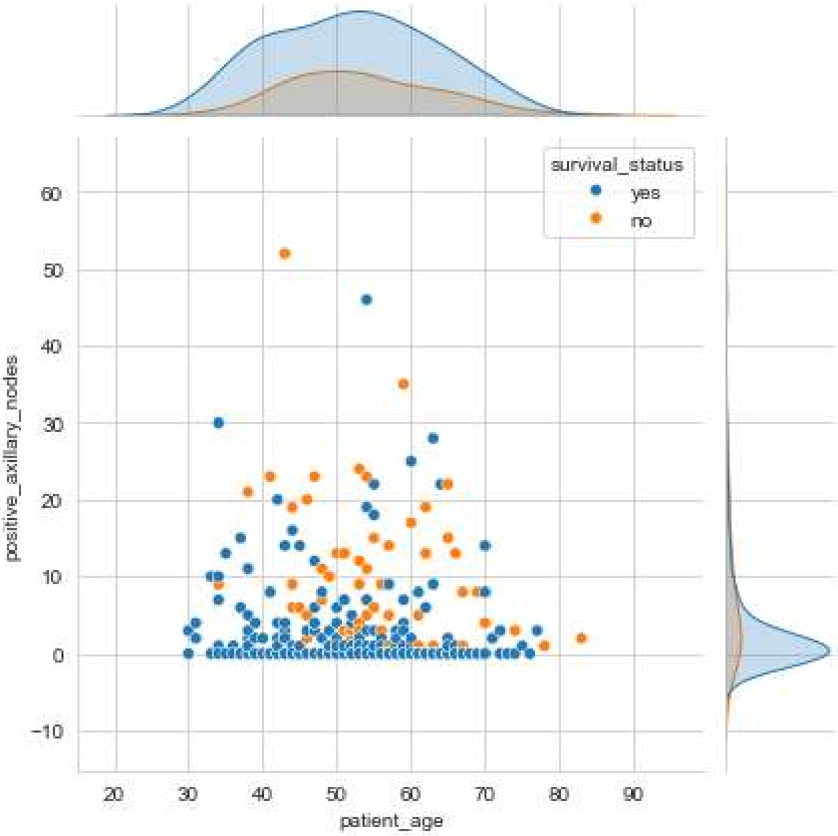


The pair plot and the joint plot reveal that there is no correlation between the patient’s age and the number of positive axillary nodes detected.

The histogram on the top edge indicates that patients are more likely to get operated in the age of 40-60 years compared to other age groups.

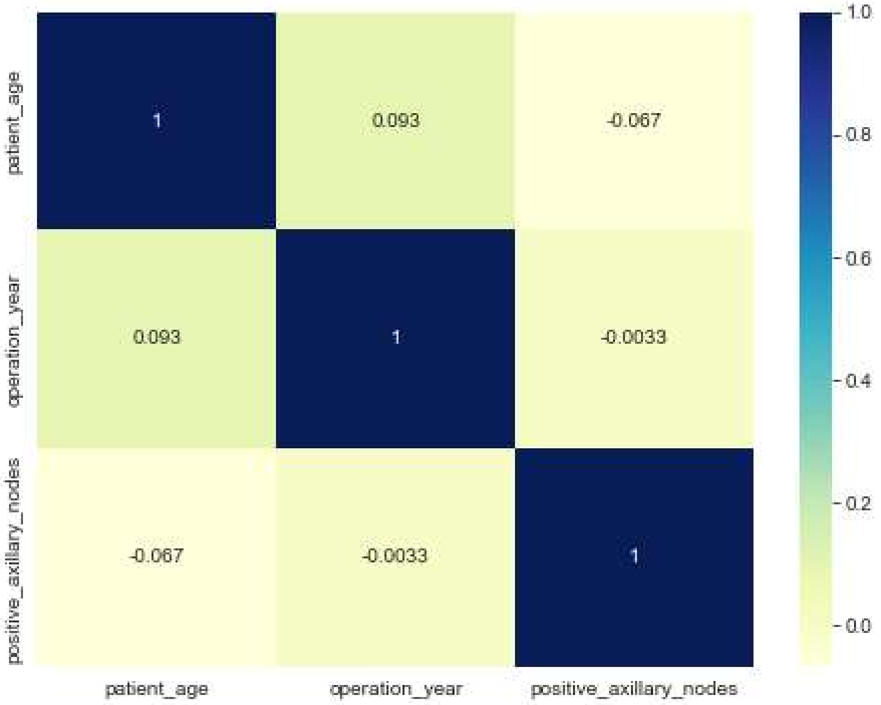
The histogram on the right edge indicates that the majority of patients had fewer than 4 positive axillary nodes.

In [18]:
```
sns.jointplot(x="patient_age",y="positive_axillary_nodes",data=df,hue="survi
plt.show()
```



Heatmap

In [19]:
```
plt.figure(figsize=(8, 6))
sns.heatmap(df.iloc[:,0:3].corr(), cmap="YlGnBu",annot=True)
plt.show()
```



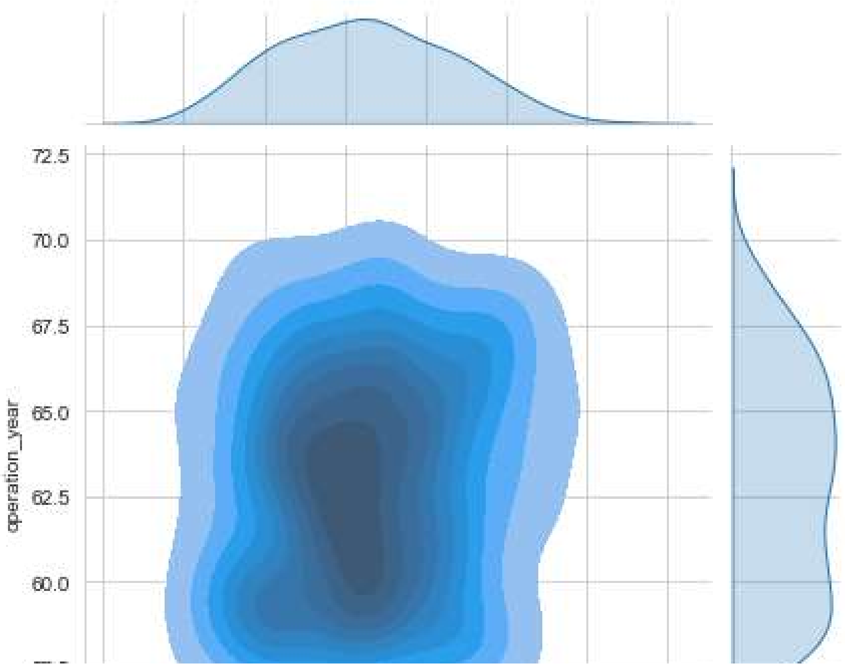
eda data analysis these values are nearly 0 for any pair, so no correlation exists among any pair of variables.

Multivariate analysis

3-dimensional surface by plotting constant z slices, called contours, in a 2-dimensional format.3d to 2d.

In [20]:

```
sns.jointplot(x = 'patient_age',  y = 'operation_year' , data = df,  kind =
plt.show()
```



years 1959–1964 witnessed more patients in the age group of 45–55 years