In [ ]:

```
# Dataset: Differentiated Thyroid Cancer Recurrence

This data set contains 13 clinicopathologic features aiming to predict recur

Gathering Data
```

In [1]:

```
# Importing necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

#import dataset
patient_data = pd.read_csv('https://raw.githubusercontent.com/karanzijm/MLEx
patient_data
```

Out[1]:

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	A
0	27	F	No	No	No	Euthyroid	Single nodular goiter-left	
1	34	F	No	Yes	No	Euthyroid	Multinodular goiter	
2	30	F	No	No	No	Euthyroid	Single nodular goiter-right	
3	62	F	No	No	No	Euthyroid	Single nodular goiter-right	
4	62	F	No	No	No	Euthyroid	Multinodular goiter	
...	...	...	...	...	...	...	...	
378	72	M	Yes	Yes	Yes	Euthyroid	Single nodular goiter-right	
379	81	M	Yes	No	Yes	Euthyroid	Multinodular goiter	
380	72	M	Yes	Yes	No	Euthyroid	Multinodular goiter	
381	61	M	Yes	Yes	Yes	Clinical Hyperthyroidism	Multinodular goiter	
382	67	M	Yes	No	No	Euthyroid	Multinodular goiter	

383 rows × 17 columns

In [ ]:

```
# Patient columns
-       T: Tumor size and extent
    -   T1a and T1b: Indicates a small tumor size, typically less than 2cm i
-       N: Lymph node involvement.
    - N0: No regional lymph node metastasis (cancer has not spread to nearby
    - N1b: Cancer has spread to certain lymph nodes (such as cervical or upp
-       M: Distant metastasis.
    - M0: No distant metastasis (cancer has not spread to other parts of the
-       Hx Smoking: History of smoking
-       Hx Radiotherapy: History of radiotherapy.
-       Thyroid Function: The functional state of the thyroid.
  -     Euthyroid: This means that the thyroid is functioning normally. The
  - Clinical Hyperthyroidism: This indicates that the patient has overactive
  -     Clinical Hypothyroidism: This indicates that the patient has underac
-       Physical Examination: Results of a physical examination of the thyro
  -     Single nodular goiter-left: A single nodule (enlarged portion of the
  -     Multinodular goiter: Multiple nodules are present in the thyroid gla
- Stages:
  - Stages I & II are typically early-stage cancers, with Stage II sometimes
  - Stage III often involves larger tumors or some lymph node involvement bu
  - Stage IV is advanced, with the cancer either spreading to nearby tissues
-       Adenopathy: Swelling or disease of lymph nodes.
  -     "No" indicates no adenopathy, meaning there is no lymph node involve
-       Pathology: The study of the disease, especially cancer.
-       Focality: The number of distinct tumor sites.
  -     Uni-focal: The cancer is localized to a single focus or site within
-       Risk: The level of cancer risk or recurrence.
  -     "Low" means the patient is considered at low risk for recurrence or
- Response: The clinical assessment of how well the patient’s condition resp
```

In [ ]:

```
All columns seem to have all their data consistent at first glance. There se

Check is the data has any null and duplicate values and remove them.
```

In [3]:

```
print(patient_data.isnull().sum())
```

```
Age                     0
Gender                  0
Smoking                 0
Hx Smoking              0
Hx Radiothreapy         0
Thyroid Function        0
Physical Examination    0
Adenopathy              0
Pathology               0
Focality                0
Risk                    0
T                       0
N                       0
M                       0
Stage                   0
Response                0
Recurred                0
dtype: int64
```

In [5]:

```
print(patient_data.duplicated().sum())
patient_data = patient_data.drop_duplicates()
```

```
19
```

In [7]:

```
patient_data.sample(5)
```

Out[7]:

	Age	Gender	Smoking	Hx Smoking	Hx Radiothreapy	Thyroid Function	Physical Examination	Adenopa
366	64	F	No	Yes	No	Euthyroid	Multinodular goiter	
105	42	F	No	No	No	Euthyroid	Single nodular goiter-right	
89	31	M	Yes	No	No	Euthyroid	Multinodular goiter	R
238	29	F	Yes	No	No	Euthyroid	Single nodular goiter-left	
280	37	F	No	No	No	Euthyroid	Single nodular goiter-right	

In [9]:

```
print(patient_data.dtypes)
```

```
Age                       int64
Gender                   object
Smoking                  object
Hx Smoking               object
Hx Radiothreapy          object
Thyroid Function         object
Physical Examination     object
Adenopathy               object
Pathology                object
Focality                 object
Risk                     object
T                        object
N                        object
M                        object
Stage                    object
Response                 object
Recurred                 object
dtype: object
```

In [11]:

```
patient_data.Gender.value_counts()
```

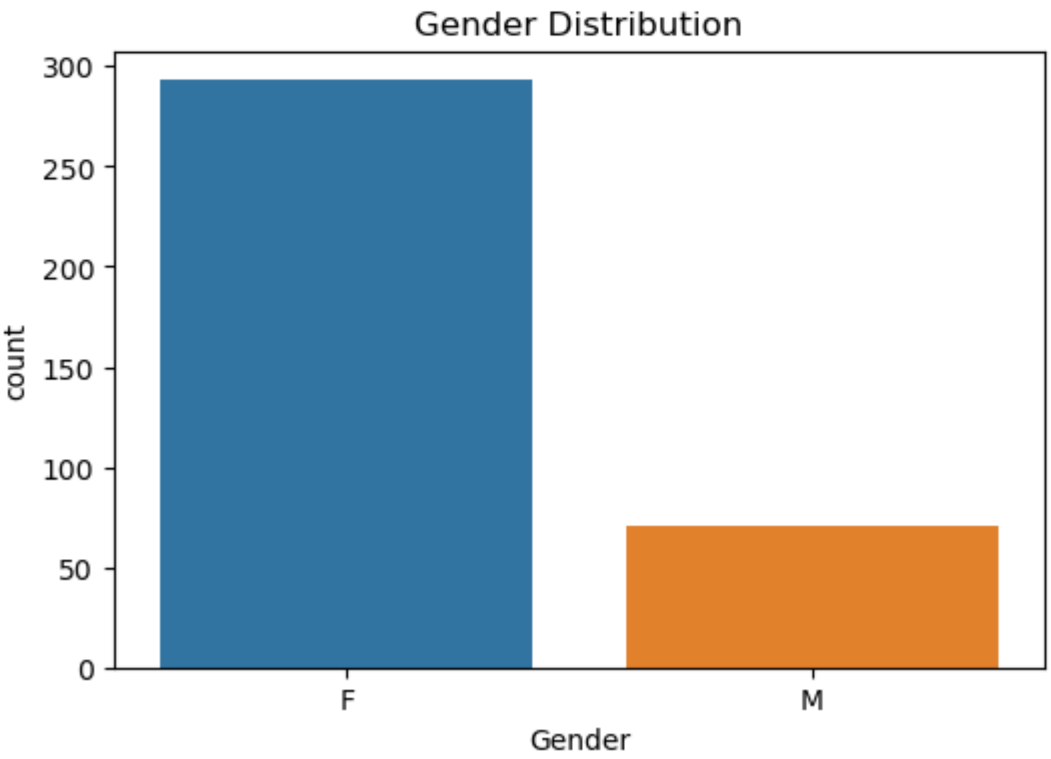
Out[11]:

```
Gender
F    293
M     71
Name: count, dtype: int64
```

In [13]:

```
plt.figure(figsize=(6, 4))
sns.countplot(x='Gender', data=patient_data)
```

```
plt.title('Gender Distribution')
plt.show()
```



In [ ]:

```
Females are more than males.
```

In [15]:

```
from scipy.stats import chi2_contingency
import numpy as np

def cramers_v(x, y):
    confusion_matrix = pd.crosstab(x, y)
    chi2 = chi2_contingency(confusion_matrix)[0]
    n = confusion_matrix.sum().sum()
    r, k = confusion_matrix.shape
    return np.sqrt(chi2 / (n * (min(k, r) - 1)))

print(cramers_v(patient_data['Hx Smoking'], patient_data['Recurred']))
print(cramers_v(patient_data['Pathology'], patient_data['Recurred']))
print(cramers_v(patient_data['Gender'], patient_data['Recurred']))
print(cramers_v(patient_data['Smoking'], patient_data['Recurred']))
print(cramers_v(patient_data['Focality'], patient_data['Recurred']))
print(cramers_v(patient_data['M'], patient_data['Recurred']))
print(cramers_v(patient_data['Stage'], patient_data['Recurred']))
print(cramers_v(patient_data['Adenopathy'], patient_data['Recurred']))
print(cramers_v(patient_data['T'], patient_data['Recurred']))
print(cramers_v(patient_data['N'], patient_data['Recurred']))
```

```
0.11718749999999997
0.25107586767432133
0.3101425326756299
0.3164898153242812
0.36236533566053664
0.3372913838647582
0.4993900250890623
0.6331196116773835
0.5996441476424428
0.624612084273687
```

In [ ]:

```
Above figures show the association of the individual features and recurrence

From the figures, Tumor size(T), Lymph node involvement(N) and Adenopathy ha
```

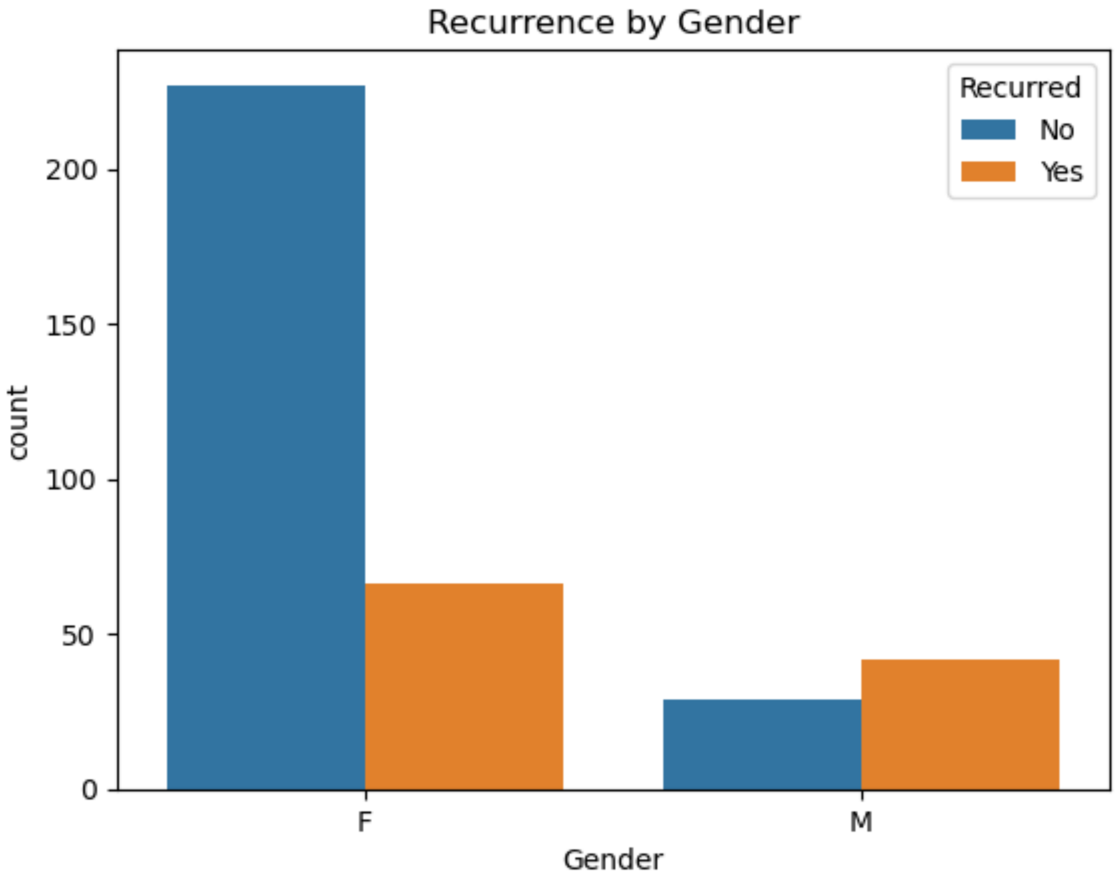
In [ ]:

```
We can also use the Chi-square & p-value as another way to determine the ext

Chi-square: This value measures the difference between the observed data and

P-value: This value indicates significance of the correlation between observ
```

In [17]:

```
ct_gender = pd.crosstab(patient_data['Gender'], patient_data['Recurred'])
print(ct_gender)

chi2, p, dof, expected = chi2_contingency(ct_gender)
print(f'Chi-square: {chi2}, p-value: {p}')

# Graph To Analyzing the effect of Gender on Recurred
sns.countplot(data=patient_data, x='Gender', hue='Recurred')
plt.title('Recurrence by Gender')
plt.show()
```

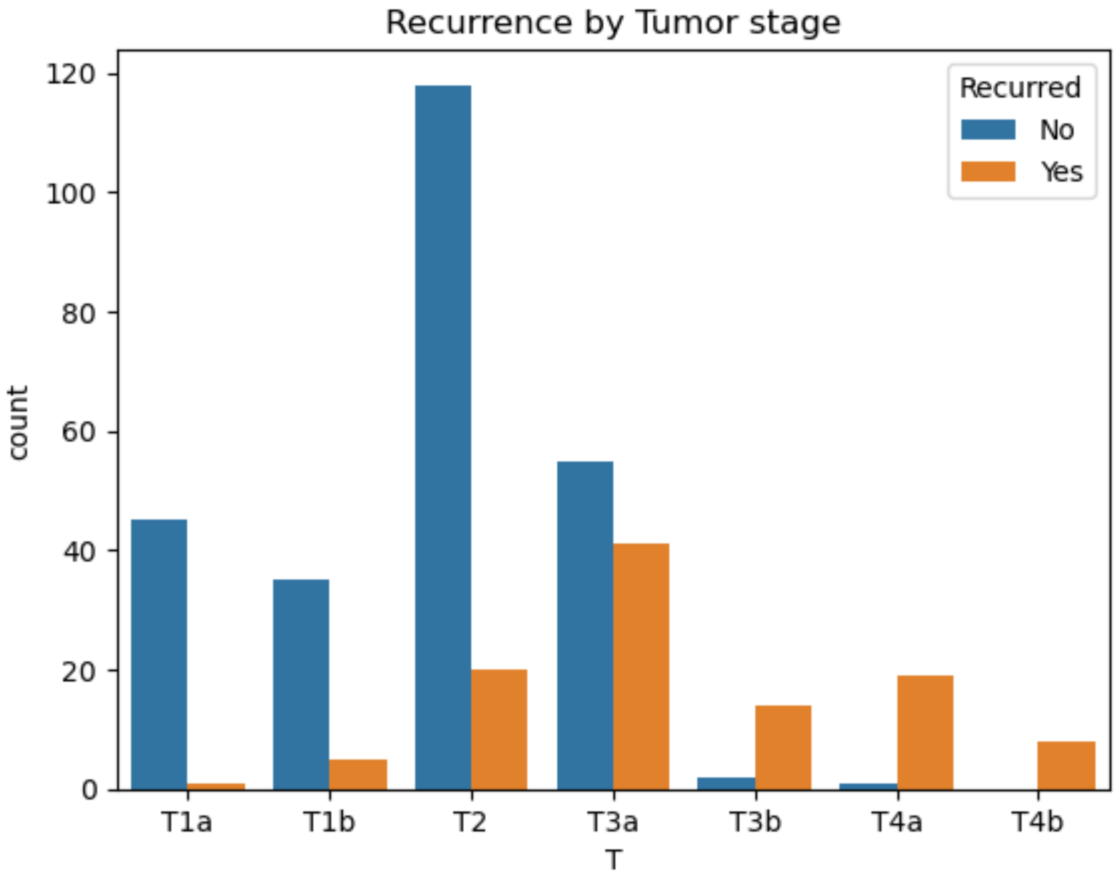
```
Recurred   No  Yes
Gender
F         227   66
M          29   42
Chi-square: 35.01257416910131, p-value: 3.2758306763157053e-09
```



```
In [ ]:  There is a statistical significant relationship between gender and recurrenc
         Males seem to have a higher recurrence rate compared to females.
         But it may have been better to have more male specimen so as to draw a bette
```

```
In [19]: ct_t = pd.crosstab(patient_data['T'], patient_data['Recurred'])
         print(ct_t)

         chi2, p, dof, expected = chi2_contingency(ct_t)
         print(f'Chi-square: {chi2}, p-value: {p}')

         # Graph To Analyzing the effect of Tumor stage on Recurred
         sns.countplot(data=patient_data, x='T', hue='Recurred')
         plt.title('Recurrence by Tumor stage')
         plt.show()
```

```
Recurred   No  Yes
T
T1a        45    1
T1b        35    5
T2        118   20
T3a        55   41
T3b         2   14
T4a         1   19
T4b         0    8
Chi-square: 130.88460978386675, p-value: 8.370007602185988e-26
```



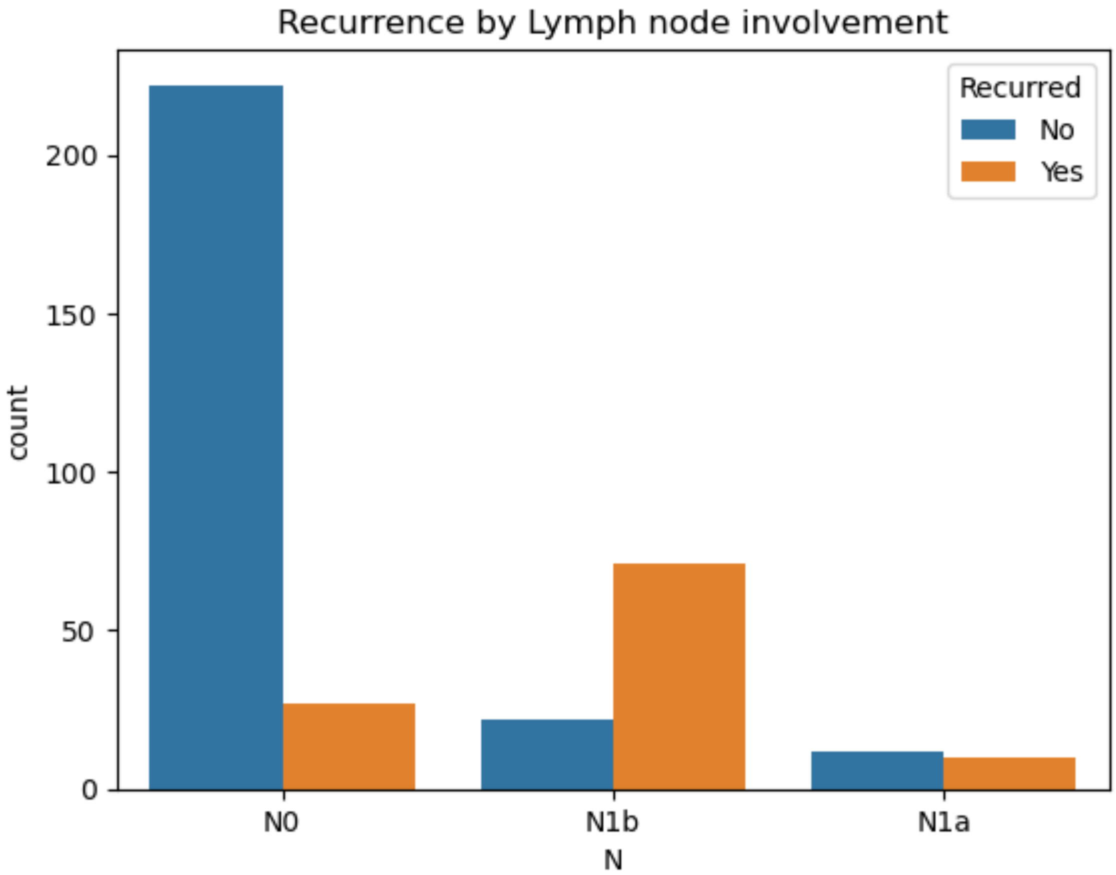
In [ ]:

```
There is a strong association between tumor stage (T) and recurrence (Recurr

- Early-stage tumors like T1a and T1b show fewer cases of recurrence.
- Later stages like T3a, T3b, T4a, and T4b show more frequent recurrence, in

This result is highly statistically significant, meaning that the stage of t
However, this test only shows association, not causation. Tumor size/stage i
```

In [21]:

```
ct_n = pd.crosstab(patient_data['N'], patient_data['Recurred'])
print(ct_n)

chi2, p, dof, expected = chi2_contingency(ct_n)
print(f'Chi-square: {chi2}, p-value: {p}')

# Graph To Analyzing the effect of Lymph node involvement on Recurred
sns.countplot(data=patient_data, x='N', hue='Recurred')
plt.title('Recurrence by Lymph node involvement')
plt.show()
```

```
Recurred   No  Yes
N
N0        222   27
N1a        12   10
N1b        22   71
Chi-square: 142.0110531187419, p-value: 1.4544260034549868e-31
```



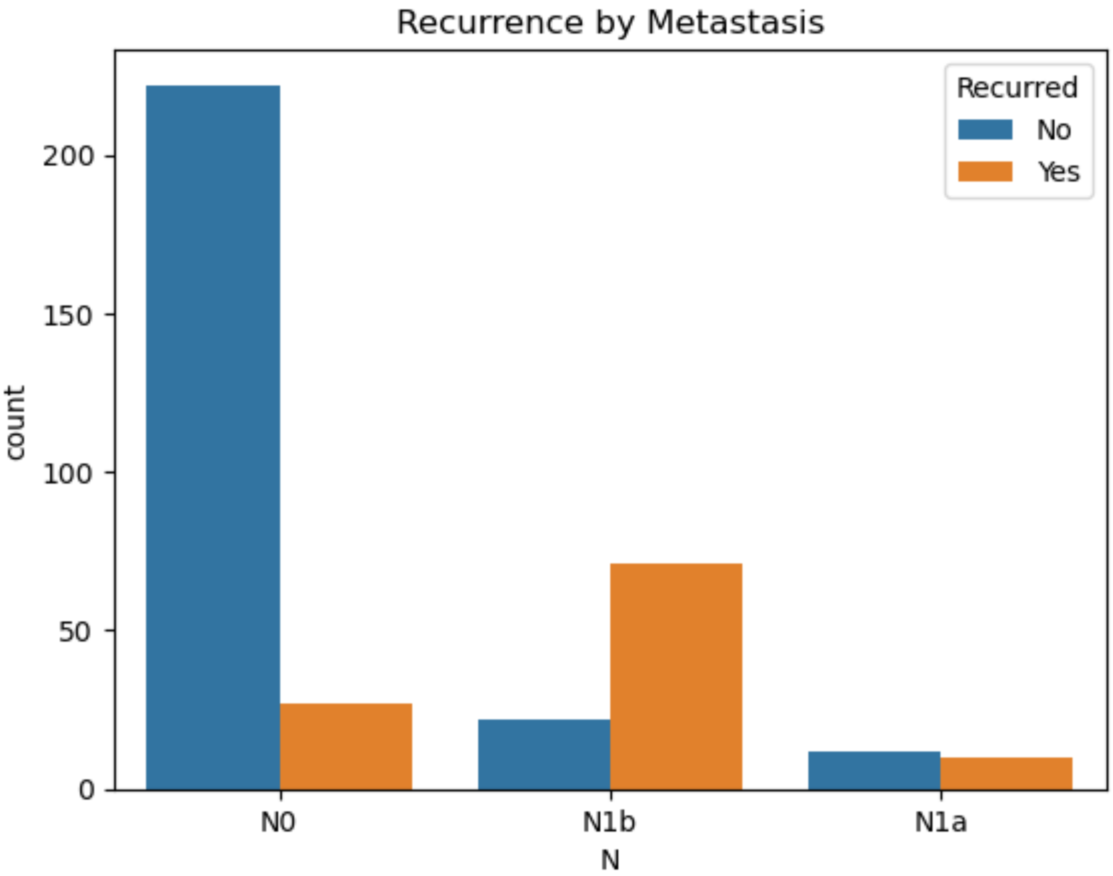
In [ ]:

```
There is a very strong and significant association between the N (lymph node
Patients with higher lymph node involvement (N1a, N1b) are much more likely
```

In [23]:

```
ct_m = pd.crosstab(patient_data['M'], patient_data['Recurred'])
print(ct_m)

chi2, p, dof, expected = chi2_contingency(ct_m)
print(f'Chi-square: {chi2}, p-value: {p}')

# Graph To Analyzing the effect of Metastasis
sns.countplot(data=patient_data, x='N', hue='Recurred')
plt.title('Recurrence by Metastasis')
plt.show()
```

```
Recurred   No  Yes
M
M0        256   90
M1          0   18
Chi-square: 41.41063385710294, p-value: 1.2338437472012865e-10
```



In [ ]:

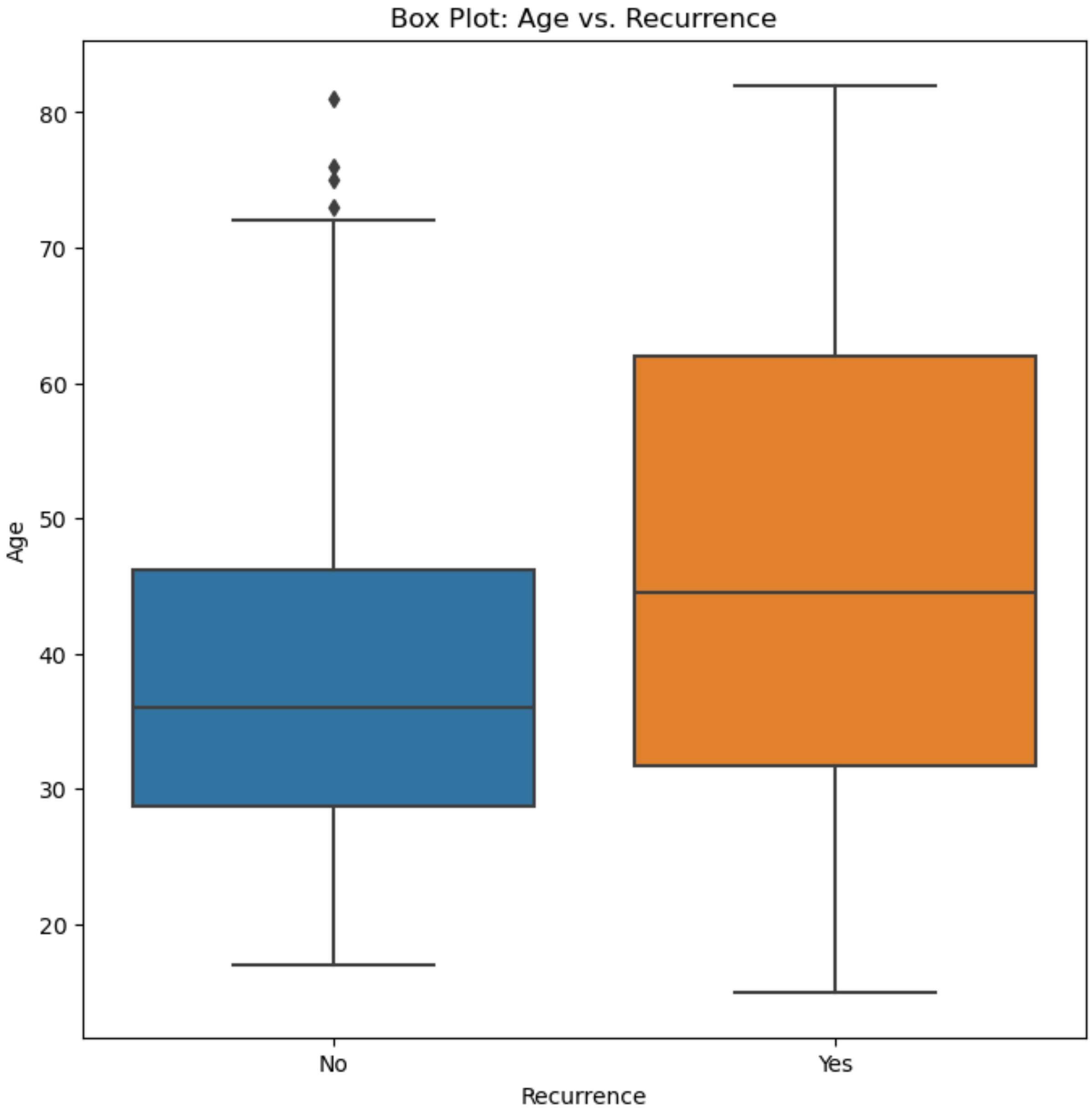
```
There is a statistically significant relationship between distant metastasis
 - M1 (metastasis present): Every patient in this group experienced recurren
 - M0 (no metastasis): Even though a portion of patients without metastasis 

The test strongly indicates that the presence of metastasis (M1) is associat

This finding suggests that metastasis is a key factor influencing the outcom
```

In [ ]:

```
All the above Chi-square values show that Tumor size(T), Lymph node involvem
```

In [27]:

```
plt.figure(figsize=(8, 8))
sns.boxplot(x='Recurred', y='Age', data=patient_data)
plt.title('Box Plot: Age vs. Recurrence')
plt.xlabel('Recurrence')
plt.ylabel('Age')
plt.show()
```



In [ ]:

```
Most of the cases without recurrence fall below 40 years of age. There are f
Most cases with recurrence are of the ages 30 and above, with the biggest nu
```