Exploratory Data Analysis (EDA)

Introduction

This dataset focuses on the recurrence of differentiated thyroid cancer (DTC), a condition where individuals who have previously been treated for thyroid cancer may get it again. Differentiated thyroid cancer is the most common type of thyroid malignancy, and understanding the factors that contribute to its recurrence is crucial for improving patient outcomes and treatment strategies.

The dataset comprises records for 382 patients, featuring 16 that could influence the chance of cancer recurrence. These features include 13 clinicopathologic variables, such as tumor size, lymph node involvement, and histopathologic subtypes, along with demographic features like age and gender. The goal of this analysis is to explore how these factors interact and contribute to the risk of recurrence, as well as to quantify the impact of each feature on the likelihood of the cancer returning.

The dataset provides a robust foundation for predictive modeling and understanding the clinical implications of recurrence risk.

Below are the feature descriptions and meaning:

1. T: Tumor size and extent
    - T1: The tumor is 2 cm or smaller in its greatest dimension.
        - T1a: The tumor is 1 cm or smaller.
        - T1b: The tumor is between 1 cm and 2 cm.
    - T2: The tumor is larger than 2 cm but not more than 4 cm, and it is confined to the thyroid gland.
    - T3: The tumor is larger than 4 cm or has minimal invasion outside the thyroid.
        - T3a: The tumor is more than 4 cm but confined to the thyroid.
        - T3b: The tumor extends outside the thyroid, but the invasion is only to nearby tissues.

  - T4: The tumor has grown beyond the thyroid gland into nearby tissues.
    - T4a: The tumor has invaded nearby structures such as muscles, trachea, or larynx.
    - T4b: The tumor has grown into major blood vessels or deeper tissues in the neck, indicating more advanced disease.

2. N: Lymph node involvement.
   - N0: No regional lymph node metastasis (cancer has not spread to nearby lymph nodes).
   - N1b: Cancer has spread to certain lymph nodes (such as cervical or upper chest).
3. M: Distant metastasis.
   - M0: No distant metastasis. The cancer has not spread to other organs or distant lymph nodes.
   - M1: Distant metastasis is present. The cancer has spread to organs such as the lungs, bones, or other distant locations outside of the thyroid gland.
4. Hx Smoking: History of smoking
5. Hx Radiotherapy: History of radiotherapy.
6. Thyroid Function: The functional state of the thyroid.
   - Euthyroid: Indicates that the thyroid is functioning normally. The patient's thyroid hormones are within the normal range, indicating no overactivity (hyperthyroidism) or underactivity (hypothyroidism).
   - Clinical Hyperthyroidism: This indicates that the patient has overactive thyroid function, where the thyroid is producing too much thyroid hormone. Common symptoms include rapid heart rate, weight loss, and nervousness.
   - Clinical Hypothyroidism: This indicates that the patient has underactive thyroid function, where the thyroid is not producing enough thyroid hormone. Symptoms might include fatigue, weight gain, and cold intolerance.
7. Physical Examination: Results of a physical examination of the thyroid.
   - Single nodular goiter-left: A single nodule (enlarged portion of the thyroid) is present on the left side.
   - Multinodular goiter: Multiple nodules are present in the thyroid gland.
8. Stages:

    - Stages I & II are typically early-stage cancers, with Stage II sometimes involving distant metastasis in younger patients.
    - Stage III often involves larger tumors or some lymph node involvement but no distant metastasis.
    - Stage IV is advanced, with the cancer either spreading to nearby tissues (IVB) or distant organs (IVC).

9. Adenopathy: Swelling or disease of lymph nodes.
    - No: indicates no adenopathy, meaning there is no lymph node involvement.
    - Yes: This signifies that adenopathy is present, indicating that the lymph nodes are enlarged, possibly due to inflammation, infection, or malignancy (spread of cancer).
    - Central: Adenopathy is located in the central (or prelaryngeal, pretracheal, and paratracheal) lymph nodes, which are situated around the thyroid gland.
    - Lateral: This refers to adenopathy in the lateral neck lymph nodes (cervical nodes), which are located on the sides of the neck.
    - Bilateral: Adenopathy is present on both sides of the neck, in both the central and lateral lymph nodes.
    - Unilateral: Adenopathy is present only on one side of the neck, either in the central or lateral lymph nodes.
10. Pathology: The study of the disease, especially cancer.
    - Papillary: This is the most common type of thyroid cancer. It grows slowly and often has a good prognosis. Papillary carcinoma frequently spreads to lymph nodes in the neck but is usually treatable.
    - Micropapillary: A variant of papillary thyroid cancer, characterized by smaller tumor sizes (less than 1 cm in diameter). It is also generally associated with a favorable outcome.
    - Follicular: This is the second most common type of thyroid cancer. It tends to spread through the bloodstream, potentially affecting lungs or bones, and can be more aggressive than papillary carcinoma.
    - Hurthle Cell: A rare form of thyroid cancer, considered more aggressive than the follicular and papillary types. It is harder to treat due to its resistance to radioactive iodine therapy.
11. Focality: The number of distinct tumor sites.
    - Uni-focal: The cancer is localized to a single focus or site within the thyroid gland.
    - Multi-Focal: There are multiple nodules or tumors in different parts of the thyroid.

12. Risk: The level of cancer risk or recurrence.
    - Low Risk: indicates the patient is considered at low risk for recurrence or aggressive cancer.
    - Intermediate Risk: Indicates patients whose cancer exhibits some concerning features, such as minimal invasion into nearby tissues, lymph node involvement, or aggressive tumor variants. There is a higher chance of recurrence compared to the low-risk group.
    - High Risk: Indicates patients have larger tumors, extensive invasion into surrounding tissues, distant metastasis, or highly aggressive cancer subtypes.

13. Response: The clinical assessment of how well the patient's condition responded to treatment.
    - Excellent Response: The patient shows no signs of residual disease after treatment.
    - Biochemical Incomplete Response: Despite no visible or structural evidence of cancer on imaging, blood tests reveal elevated thyroglobulin levels, indicating that some cancer cells may remain.
    - Structural Incomplete Response: There is visible evidence of remaining cancer, either in the thyroid bed (where the thyroid was removed) or in distant sites (metastasis).
    - Indeterminate Response: The response is unclear, and there might be some changes or signals on imaging or blood tests, but it's uncertain whether these indicate active cancer or benign changes.

Analysis of Data

The data does not have any null values but has 19 duplicate values, which have been removed. Average age of the patients is 40. The youngest patient is 15 and oldest is 82.

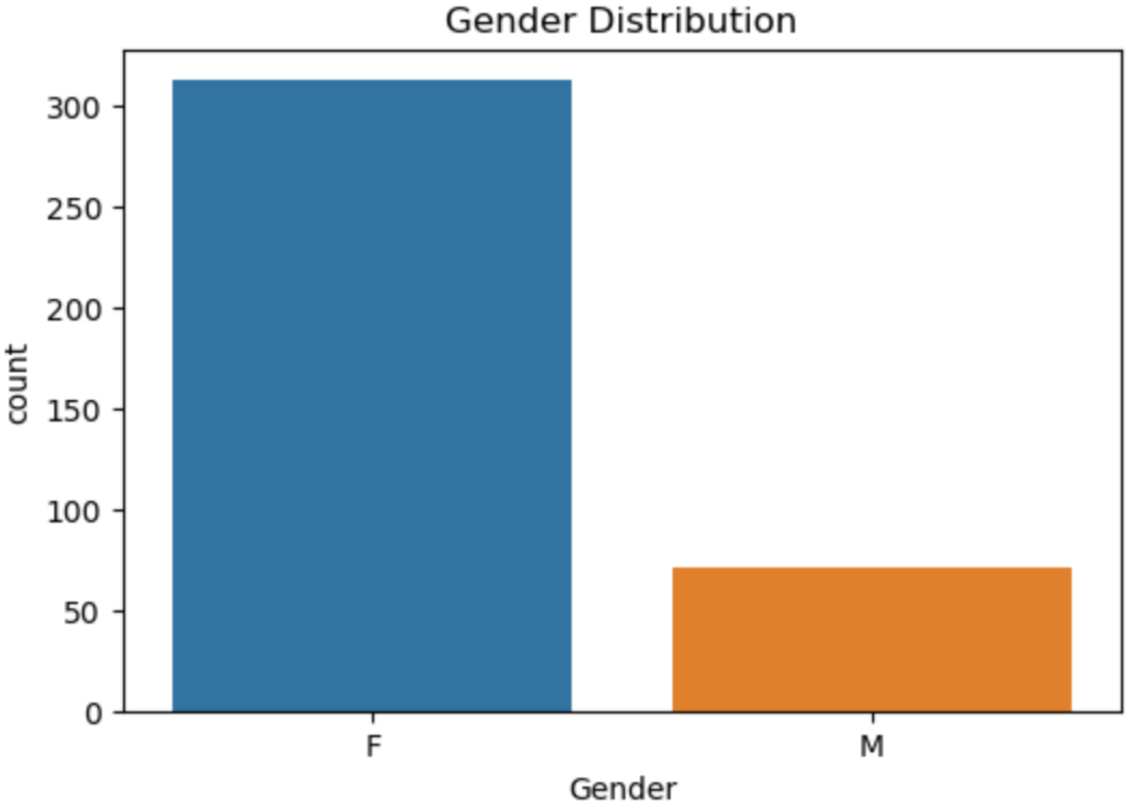


Figure 1: Gender Distribution

The graph shows that the data has more female patients than male.

We can use the Chi-square & p-value as way to determine the extent of impact the features have on recurrence.

Chi-square: This value measures the difference between the observed data and the expected values. If there is no association between a particular observed and the outcome. A high chi-square value suggests that the observed feature distribution is significantly different from what is expected under the assumption of no association.

P-value: This value indicates significance of the correlation between observed data and expected value. If it is a high value means the association between the variables is not due to chance.

Recurrence by Gender

From the data there is a statistical significant relationship between gender and recurrence. 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 better conclusion.

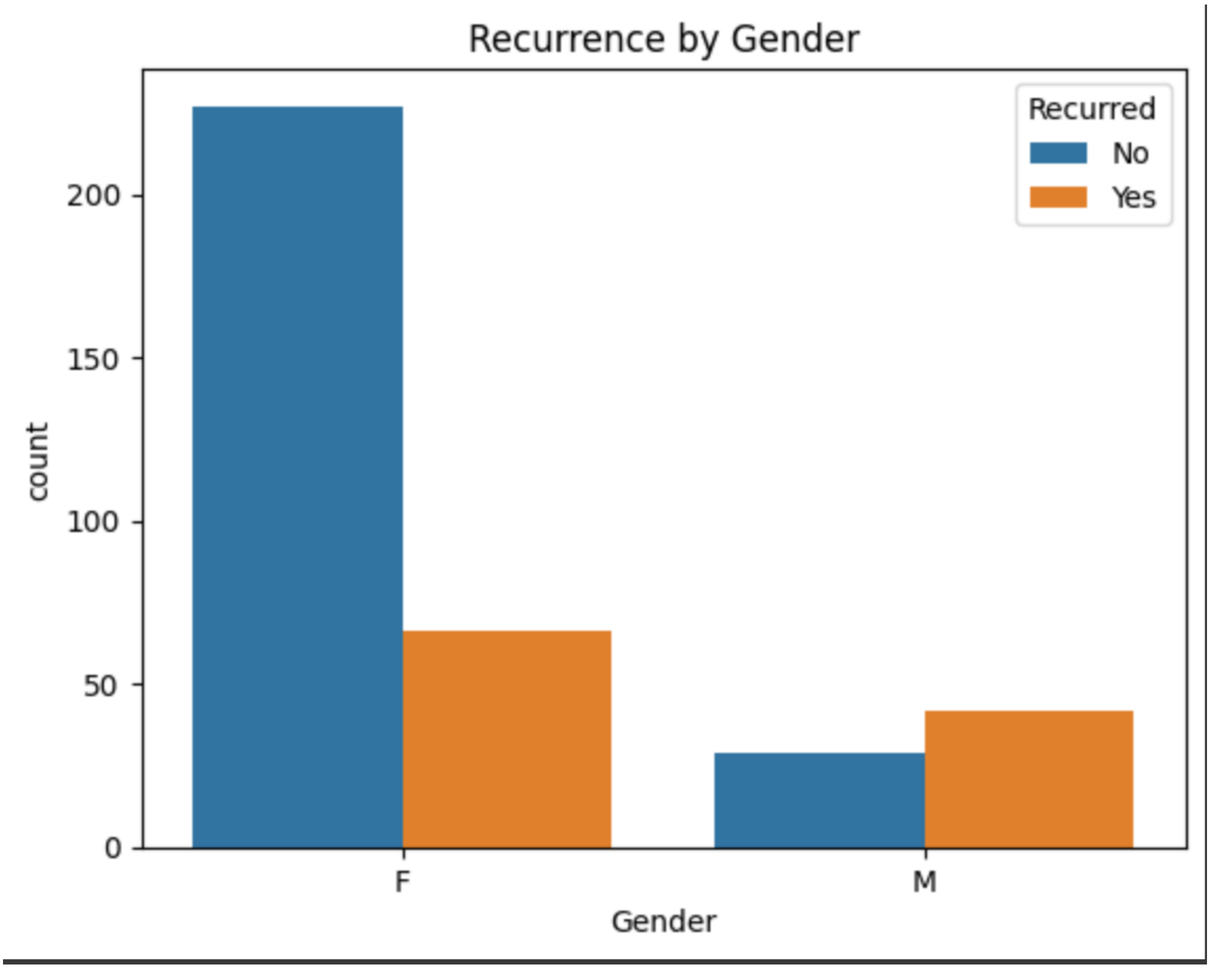


Figure 2: Recurrence by Gender

Recurrence by Tumor stage

As shown below, there is a strong association between tumor stage (T) and recurrence. As tumor stages progress from T1a to T4b, the likelihood of recurrence appears to increase as Figure 3 below shows. For example:

- Early-stage tumors like T1a and T1b show fewer cases of recurrence.

- Later stages like T3a, T3b, T4a, and T4b show more frequent recurrence, indicating that patients with more advanced tumors are significantly more likely to experience a recurrence.

This result is highly statistically significant, meaning that the stage of the tumor is very likely to impact the recurrence rate. However, this test only shows association, not causation. Tumor size/stage is strongly correlated with recurrence, but further analysis would be needed to understand the underlying reasons for this relationship or whether it and another feature(s) would show concrete significant causation for recurrence.

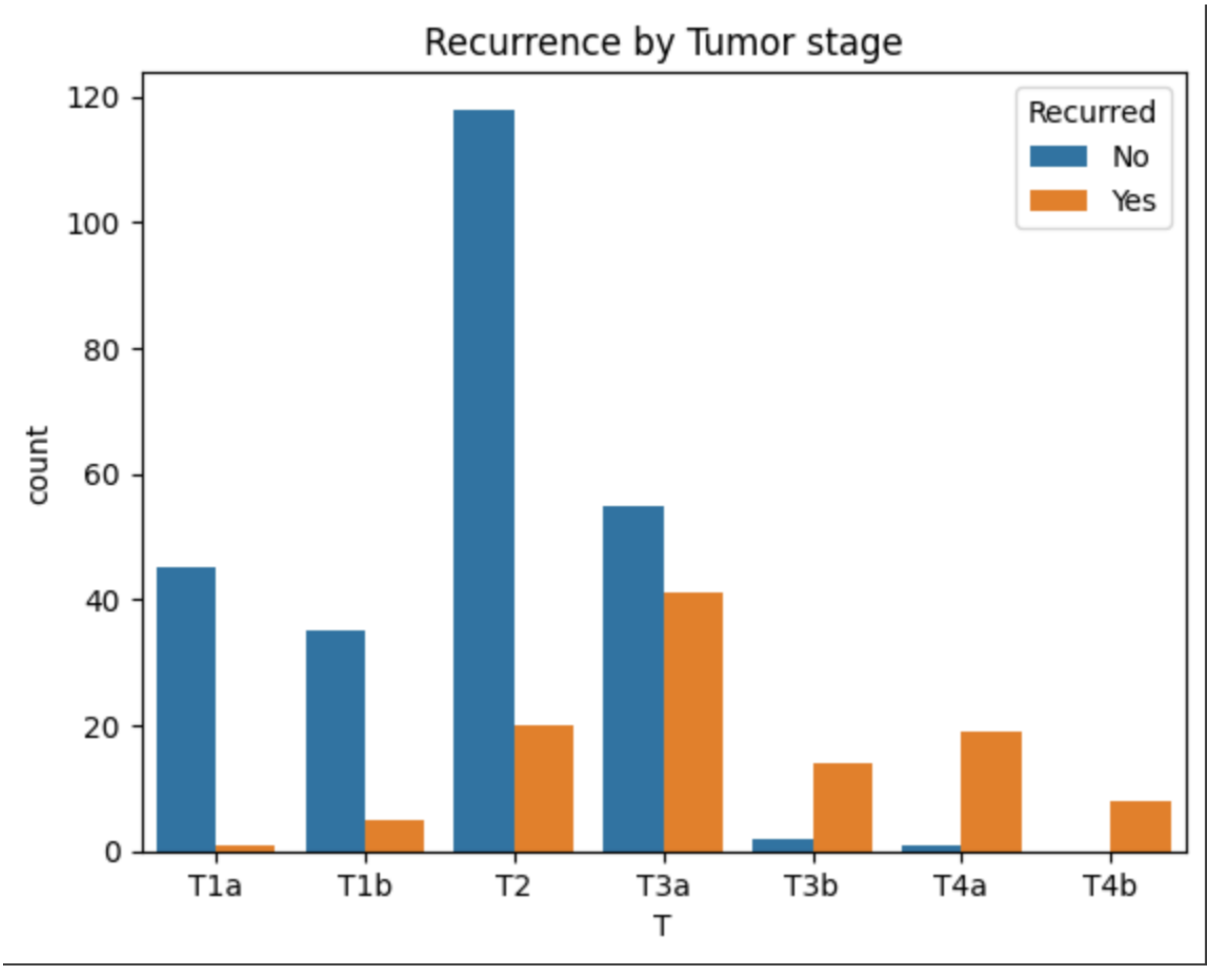


Figure 3: Reccurence by Tumor stage

Recurrence by Metastasis

There is a statistically significant relationship between distant metastasis (M) and recurrence. Specifically:

- M1 (metastasis present): Every patient in this group experienced recurrence, suggesting a very high likelihood of recurrence when metastasis is present.
- M0 (no metastasis): Even though a portion of patients without metastasis experienced recurrence, a large number (256 patients) did not.

The test strongly indicates that the presence of metastasis (M1) is associated with a higher likelihood of recurrence, and the likelihood that this association is due to chance is extremely low.

This finding suggests that metastasis is a key factor influencing the outcome of recurrence.

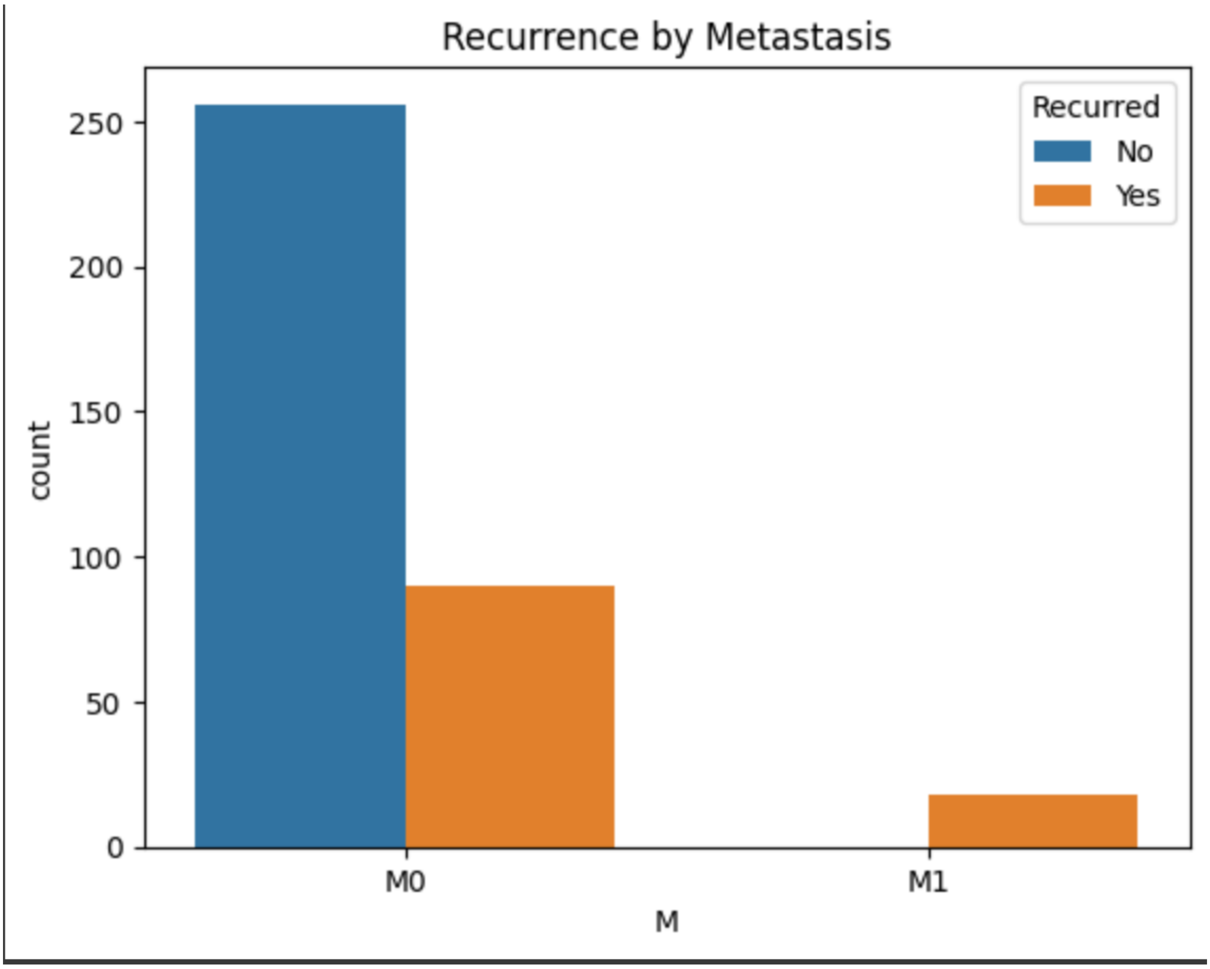
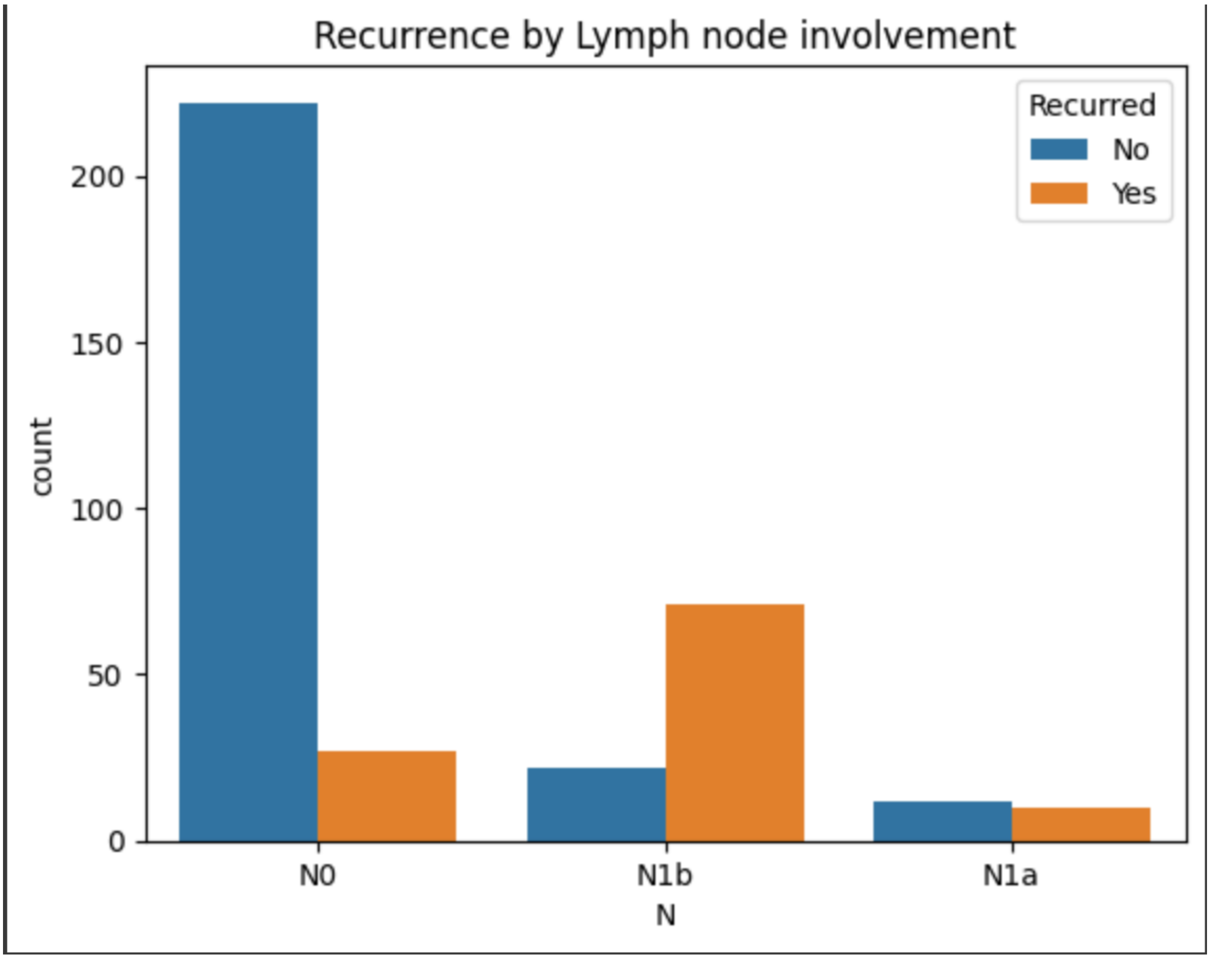


Figure 4: Reccurence by Metastasis

Reccurence by Lymph node involvement

There is a very strong and significant association between the N (lymph node involvement) variable and the likelihood of recurrence. Patients with higher lymph node involvement (N1a, N1b) are much more likely to experience recurrence than those with no lymph node involvement (N0).



All the above information shows that Tumor size(T), Lymph node involvement(N) and Adenopathy have high impact on recurrence.

Recurrence by Age

For recurrence by age we shall use a box plot to represent the correlation. Generally from the graph it shows there high number of recurrence with higher age for ages 30 and above.

The graph shows that the median age of those who do not get recurrence is around 37 years. The biggest number of those who don’t get recurrence is between the medium 30, and 50. This means that 75%, of those who don’t get recurrence lay within this range. It also shows some outliers beyond the age of 70.

For the patients that had recurrence, their median age is around 40 years. And the biggest number of them, 75%, lay between median and 70 years of age.

Age may not be a good indicator as those we've described above because for example by looking at the graph one can't tell for sure whether someone between 30 and 50 would likely have a recurrence or not. Which means that age would have to analysed along with one or more other features to get definitive possibiliry of recurrence.

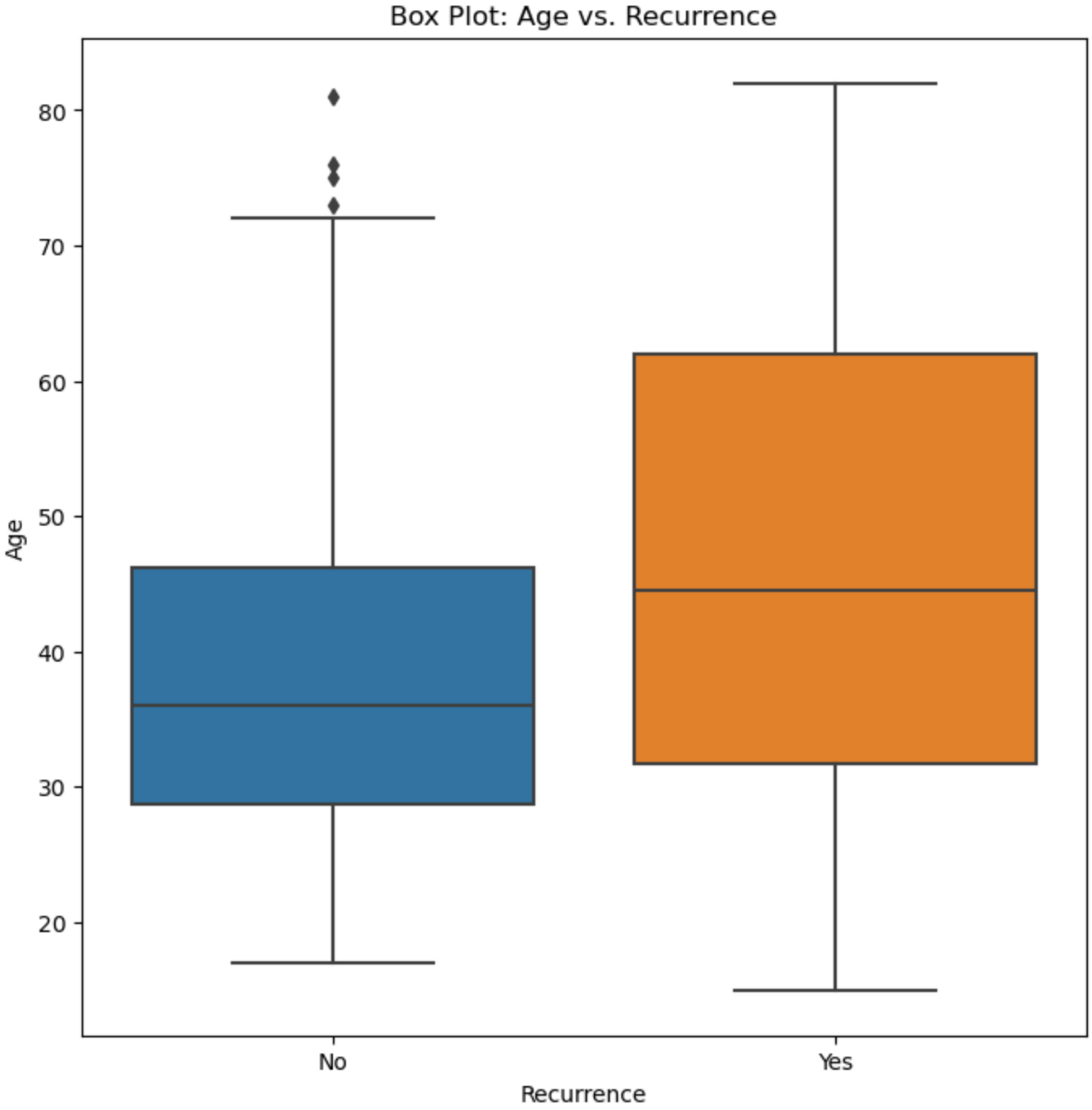


Figure 5: Reccurence by age