**Thyroid Disease Data**

<https://www.kaggle.com/datasets/jainaru/thyroid-disease-data>

This data set aiming to predict recurrence of well differentiated thyroid cancer. The data set was collected in duration of 15 years and each patient was followed for at least 10 years.

**Source**

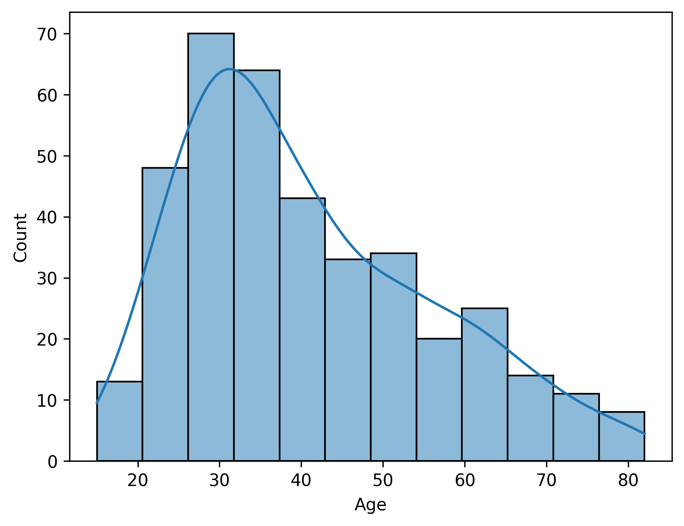
The data was procured from [thyroid disease datasets](http://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence) provided by the UCI Machine Learning Repository.

**Exploratory Data Analysis (EDA)**

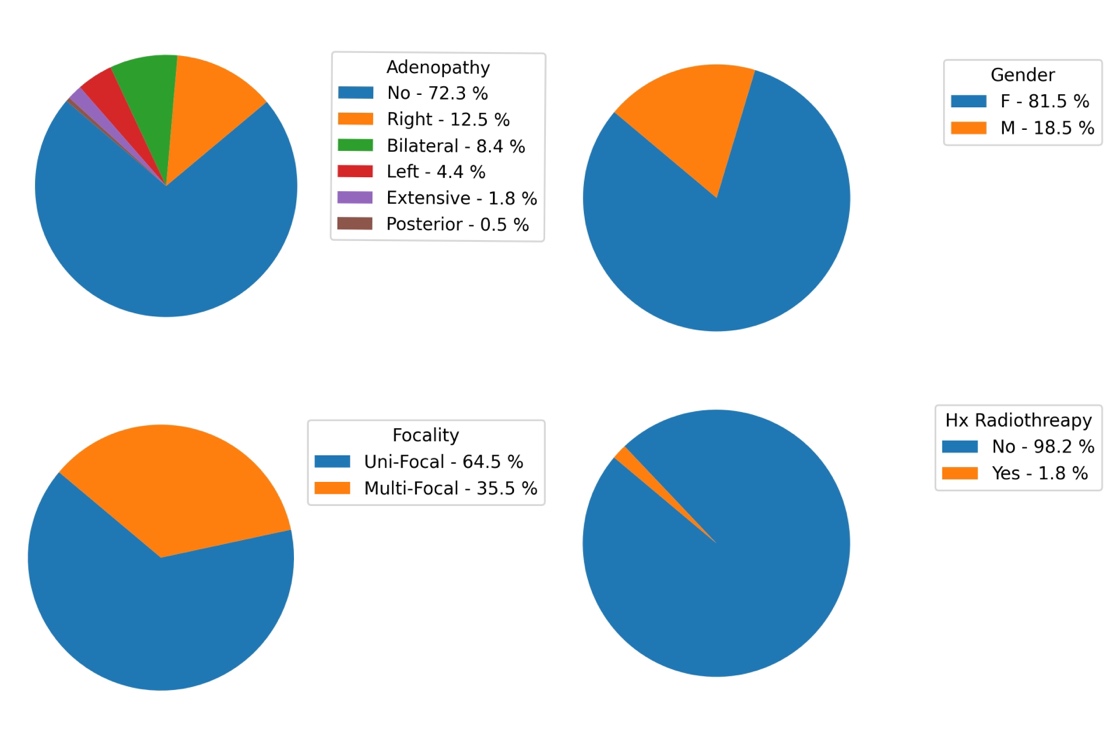
1. Response Variable

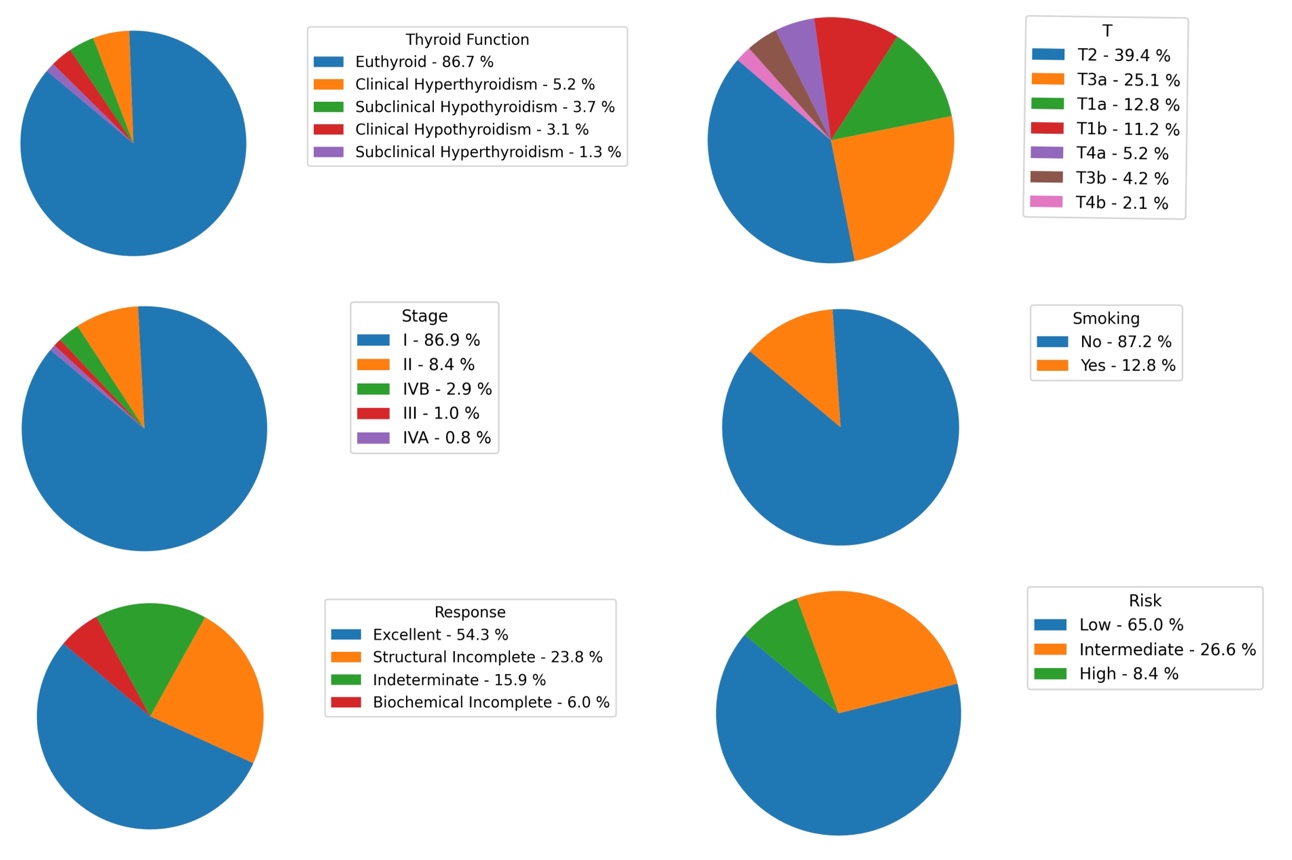
Recurred: Indicates whether the cancer has recurred after initial treatment.

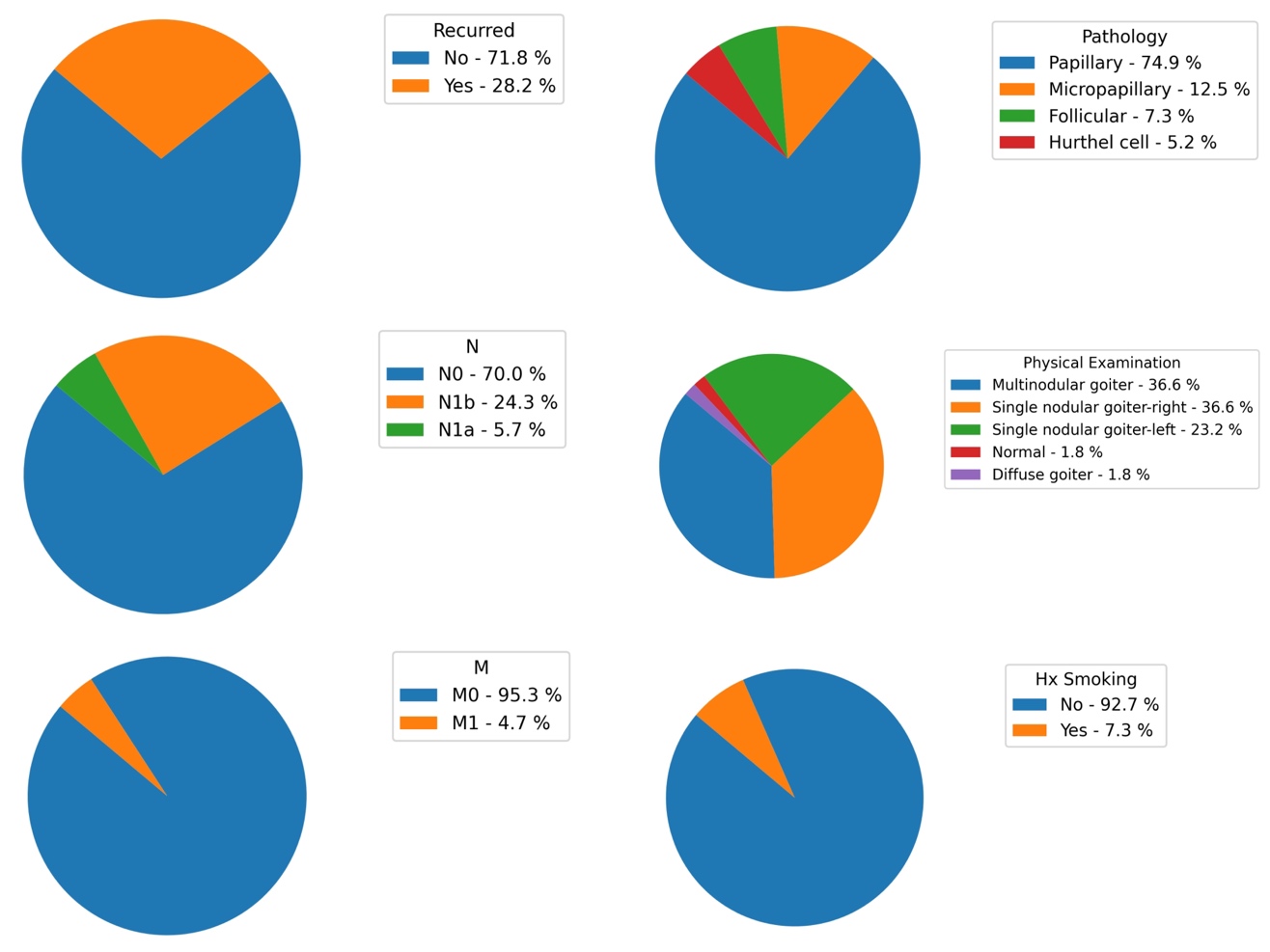
1. Explanatory Variables
2. Age: The age of the patient at the time of diagnosis or treatment.
3. Gender: The gender of the patient (male or female).
4. Smoking: Whether the patient is a smoker or not.
5. Hx Smoking: Smoking history of the patient (e.g., whether they have ever smoked).
6. Hx Radiotherapy: History of radiotherapy treatment for any condition.
7. Thyroid Function: The status of thyroid function, possibly indicating if there are any abnormalities.
8. Physical Examination: Findings from a physical examination of the patient, which may include palpation of the thyroid gland and surrounding structures.
9. Adenopathy: Presence or absence of enlarged lymph nodes (adenopathy) in the neck region.
10. Pathology: Specific types of thyroid cancer as determined by pathology examination of biopsy samples.
11. Focality: Whether the cancer is unifocal (limited to one location) or multifocal (present in multiple locations).
12. Risk: The risk category of the cancer based on various factors, such as tumor size, extent of spread, and histological type.
13. T: Tumor classification based on its size and extent of invasion into nearby structures.
14. N: Nodal classification indicating the involvement of lymph nodes.
15. M: Metastasis classification indicating the presence or absence of distant metastases.
16. Stage: The overall stage of the cancer, typically determined by combining T, N, and M classifications.
17. Response: Response to treatment, indicating whether the cancer responded positively, negatively, or remained stable after treatment.
18. Visualization by Variables:
19. Quantitative Variables



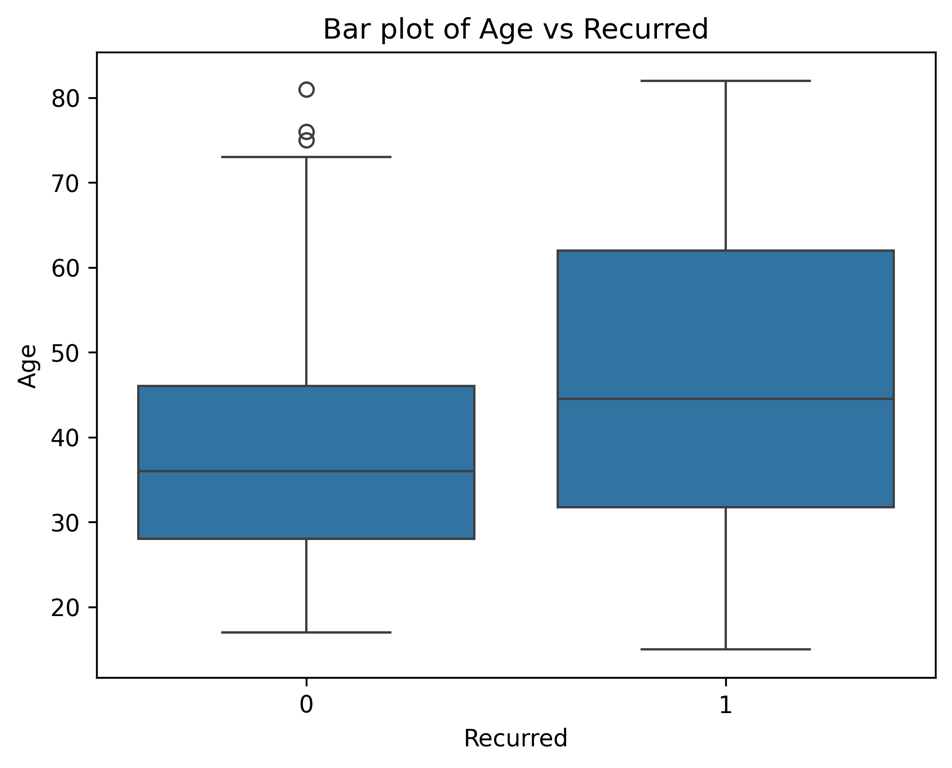
1. Categorical Variables



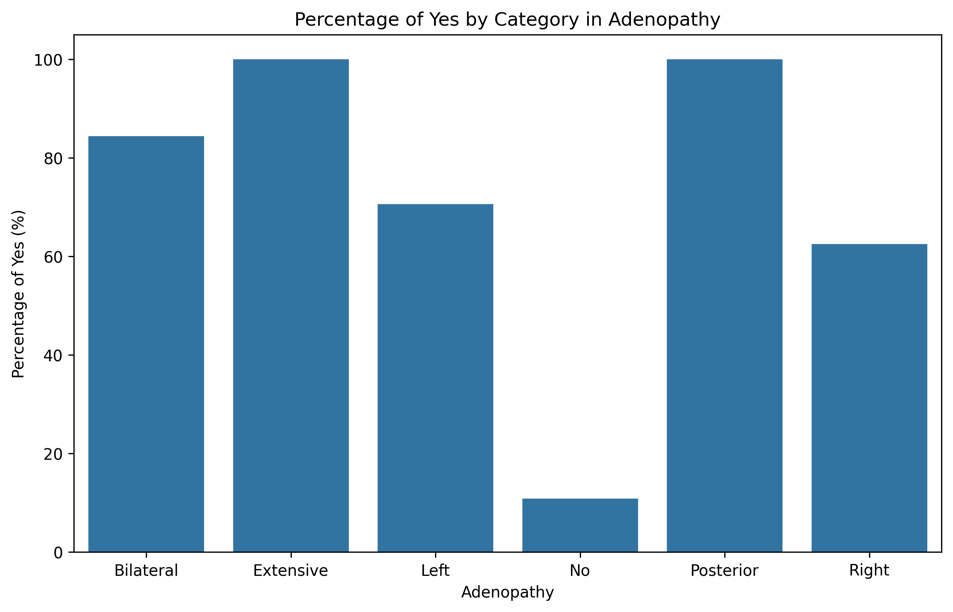


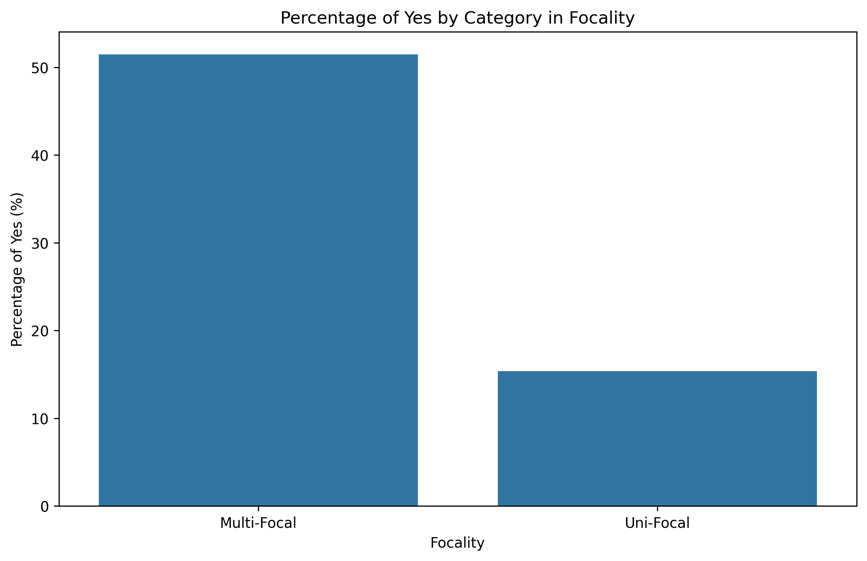


1. Potential Relationship between Explanatory Variables ad Response Variables
2. Quantitative Variables

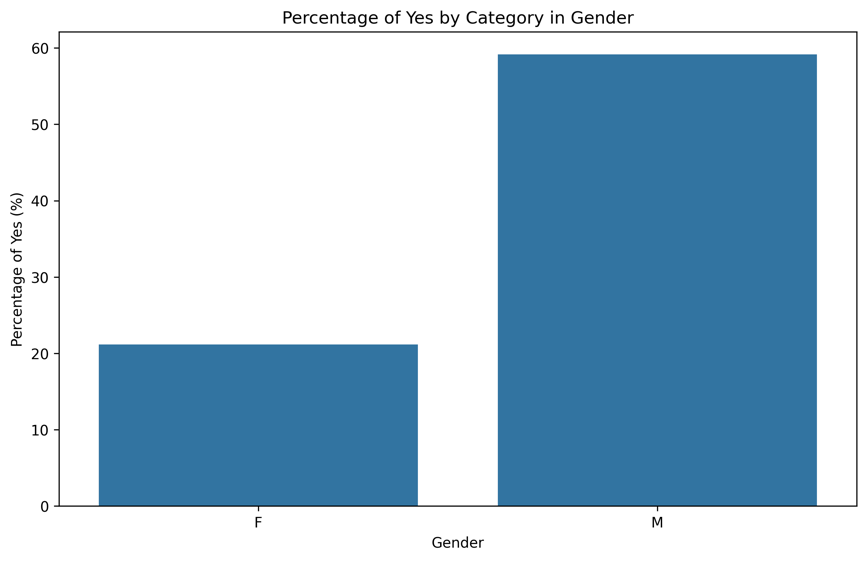


* + Older Age may lead to higher recurrence probability.

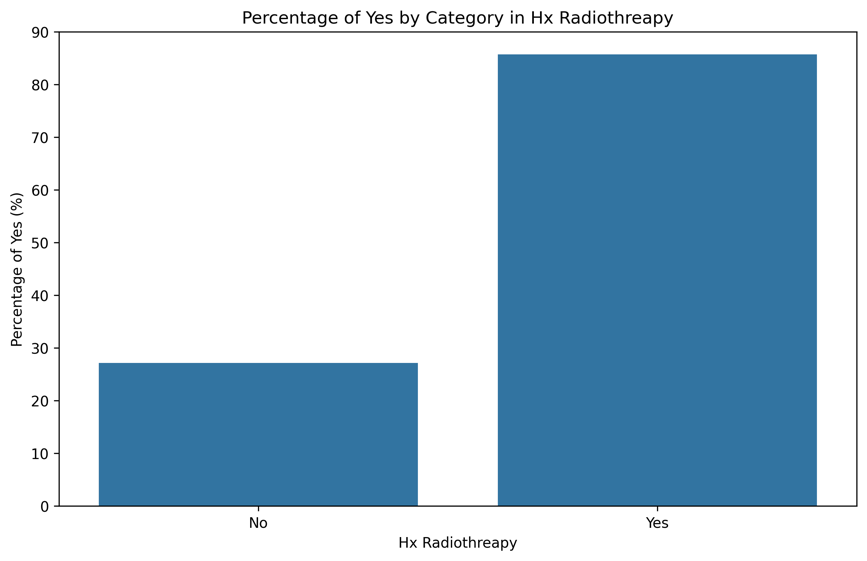
1. Categorical Variable
   * Having of enlarged lymph nodes in neck region may lead to higher recurrence probability.



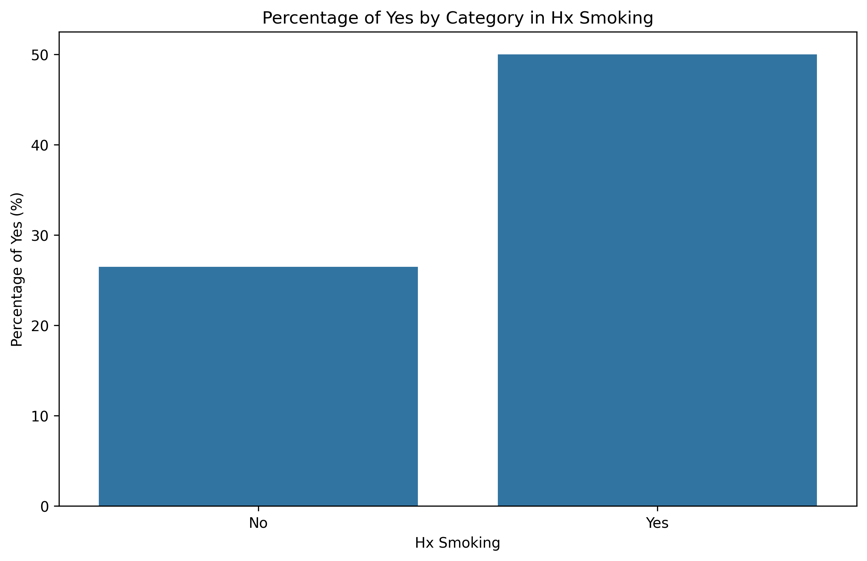
* + The cancer appeared in multiple location may have higher recurrence probability.



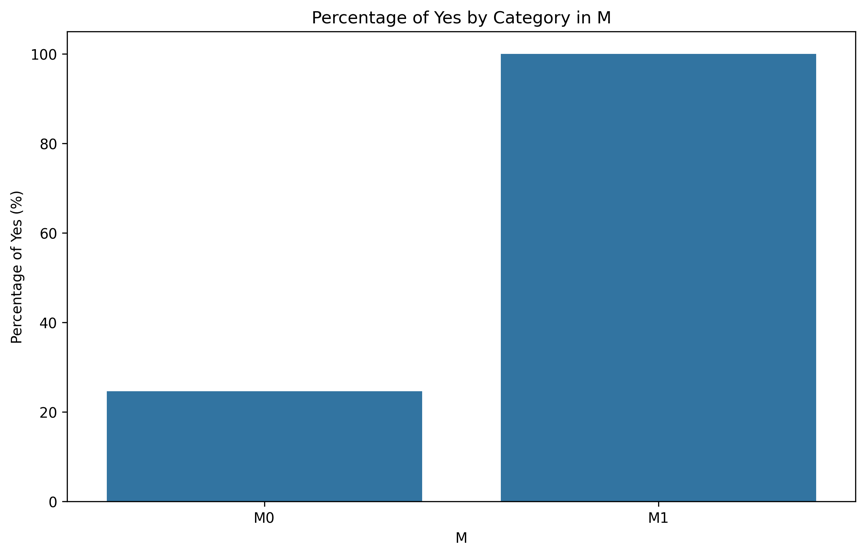
* + Males may have a higher recurrence probability.



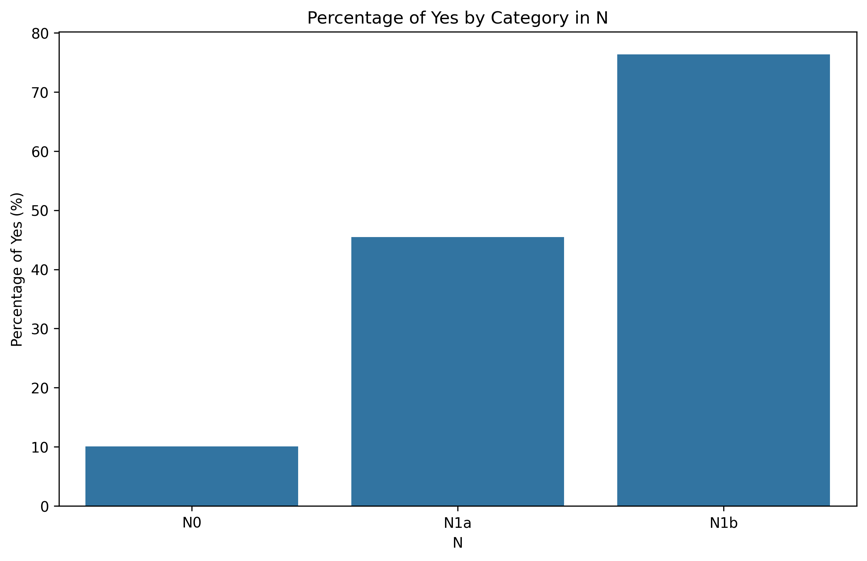
* + Previous Radio treatment may lead to a higher recurrence probability.



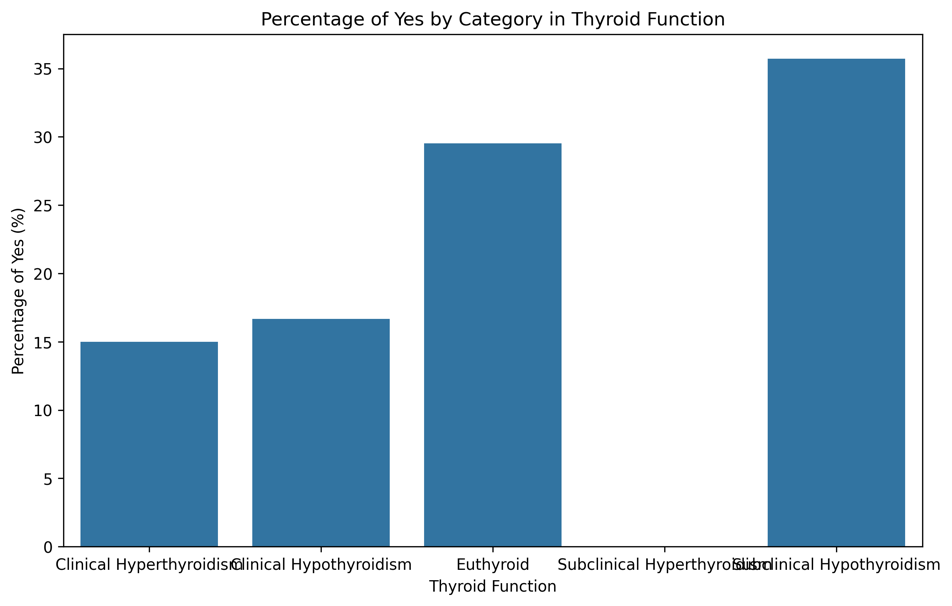
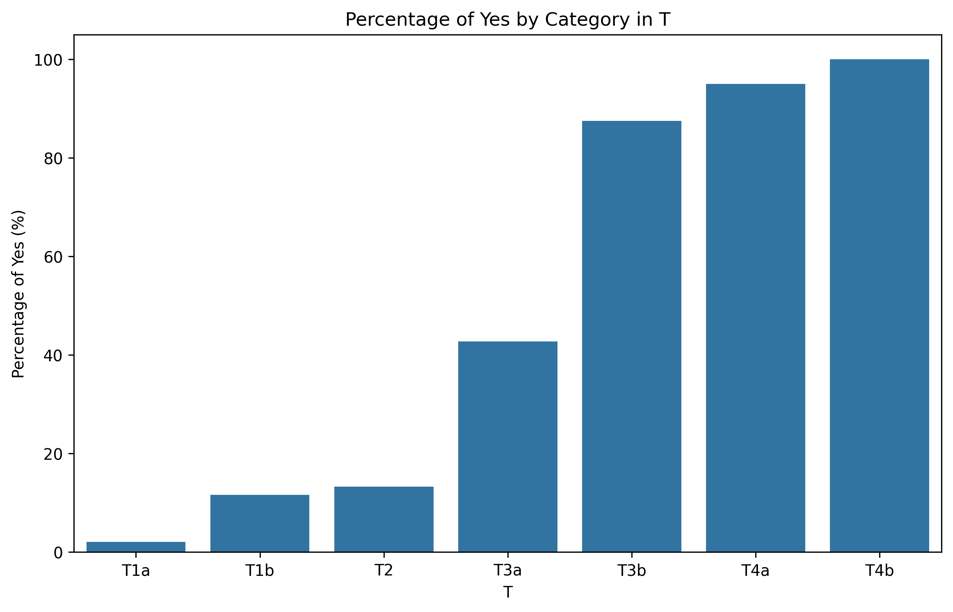
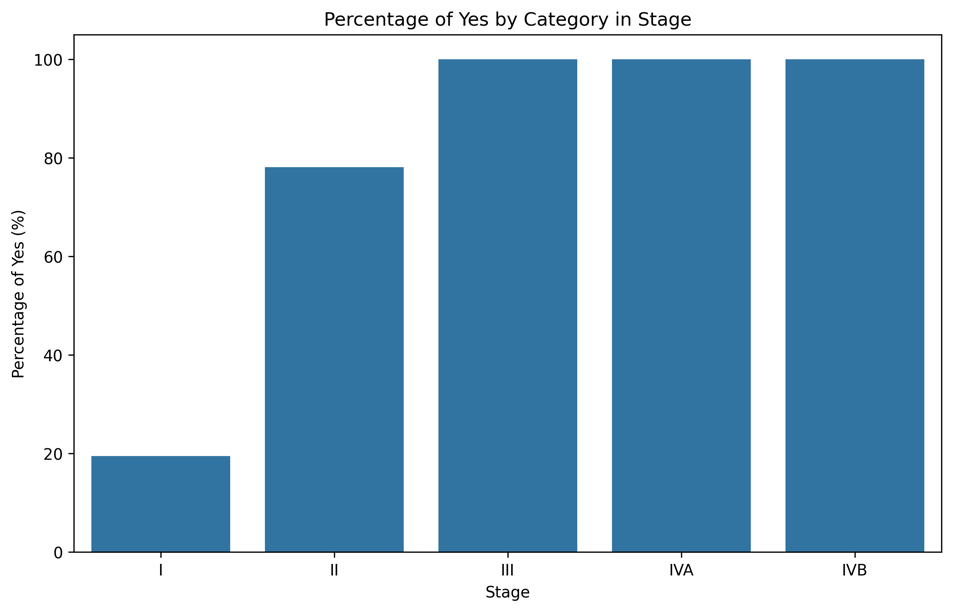
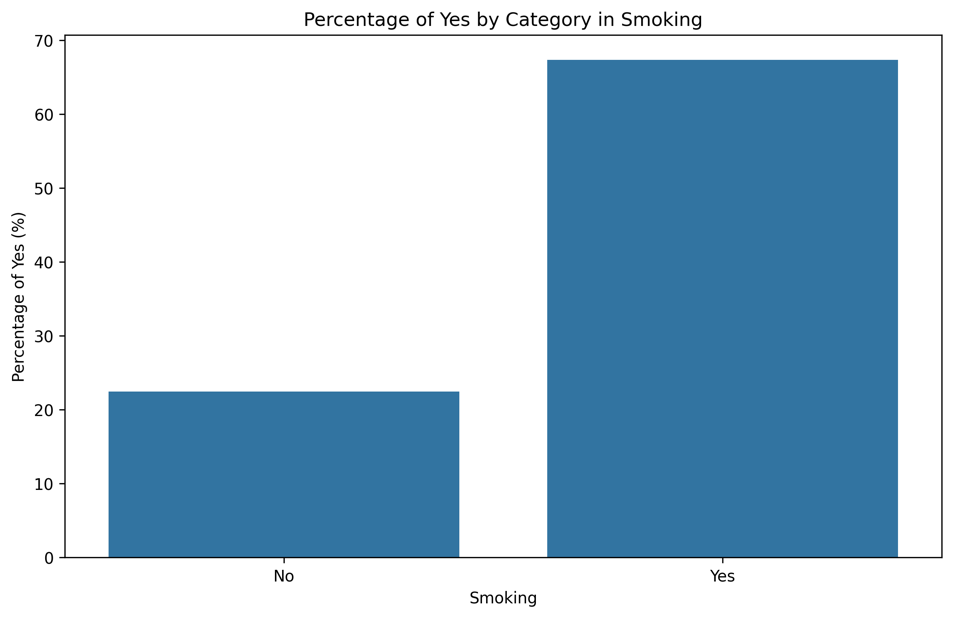
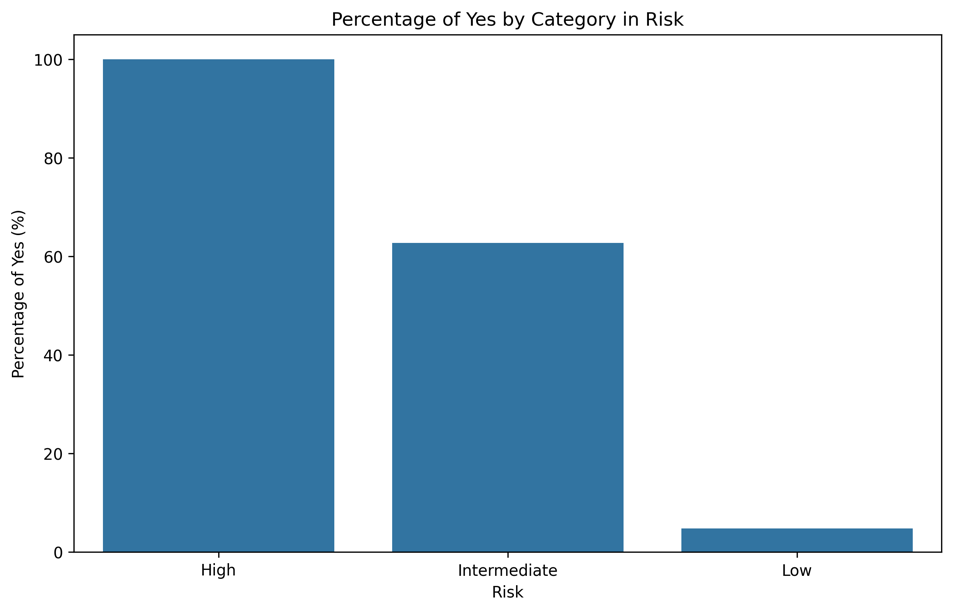
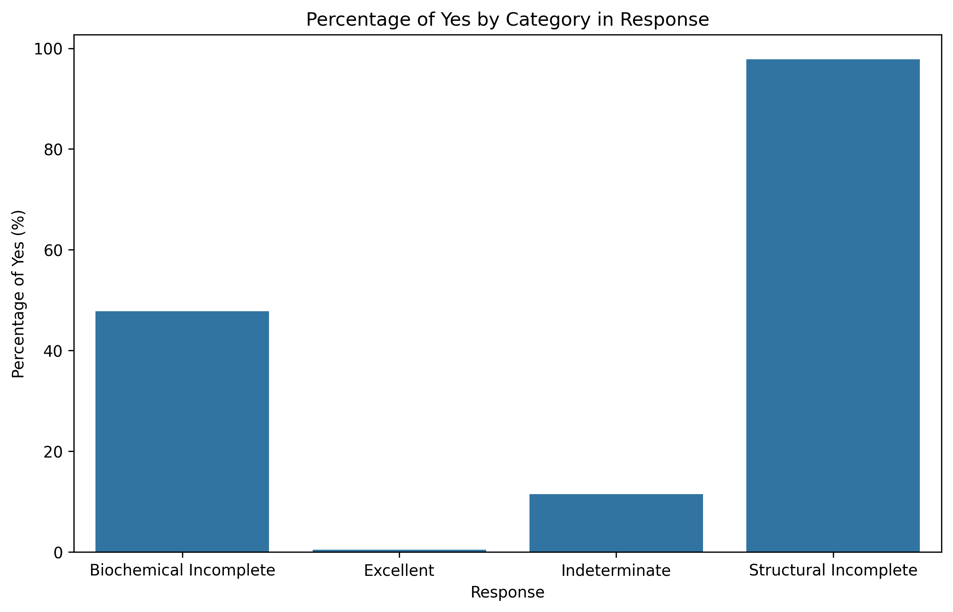
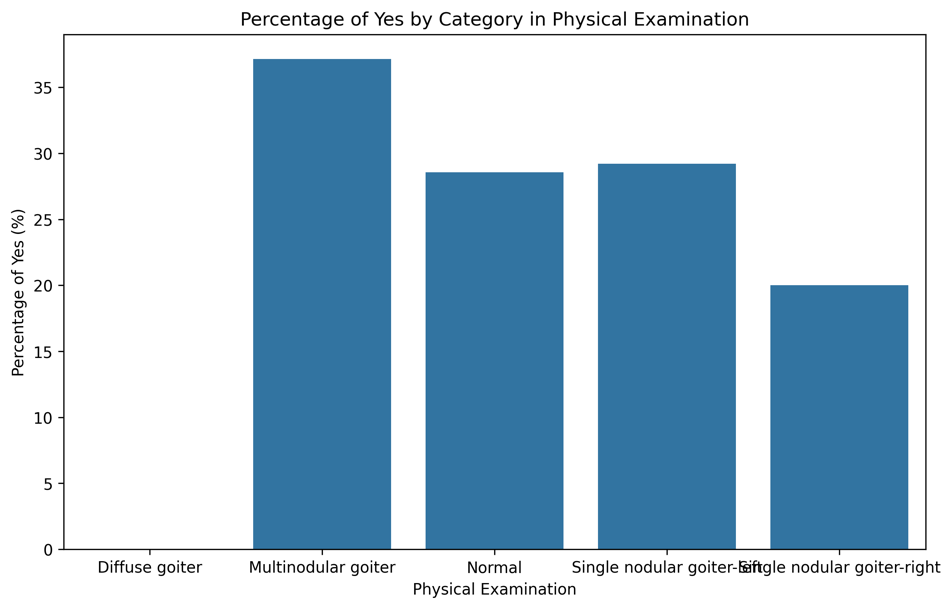
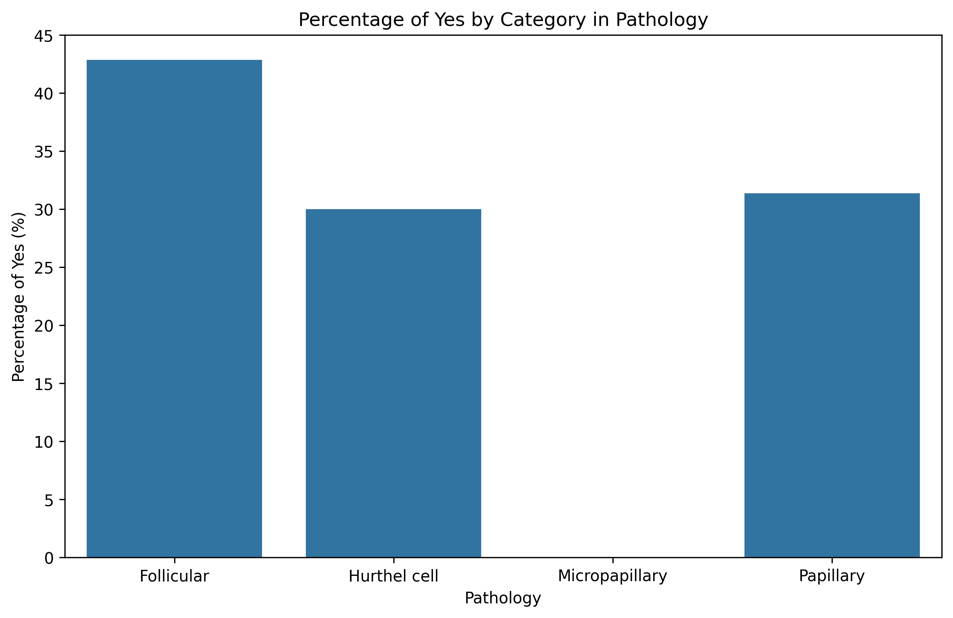
* + Previous Smoking History may lead to a higher recurrence probability



* + Distant metastases may lead to a higher recurrence probability

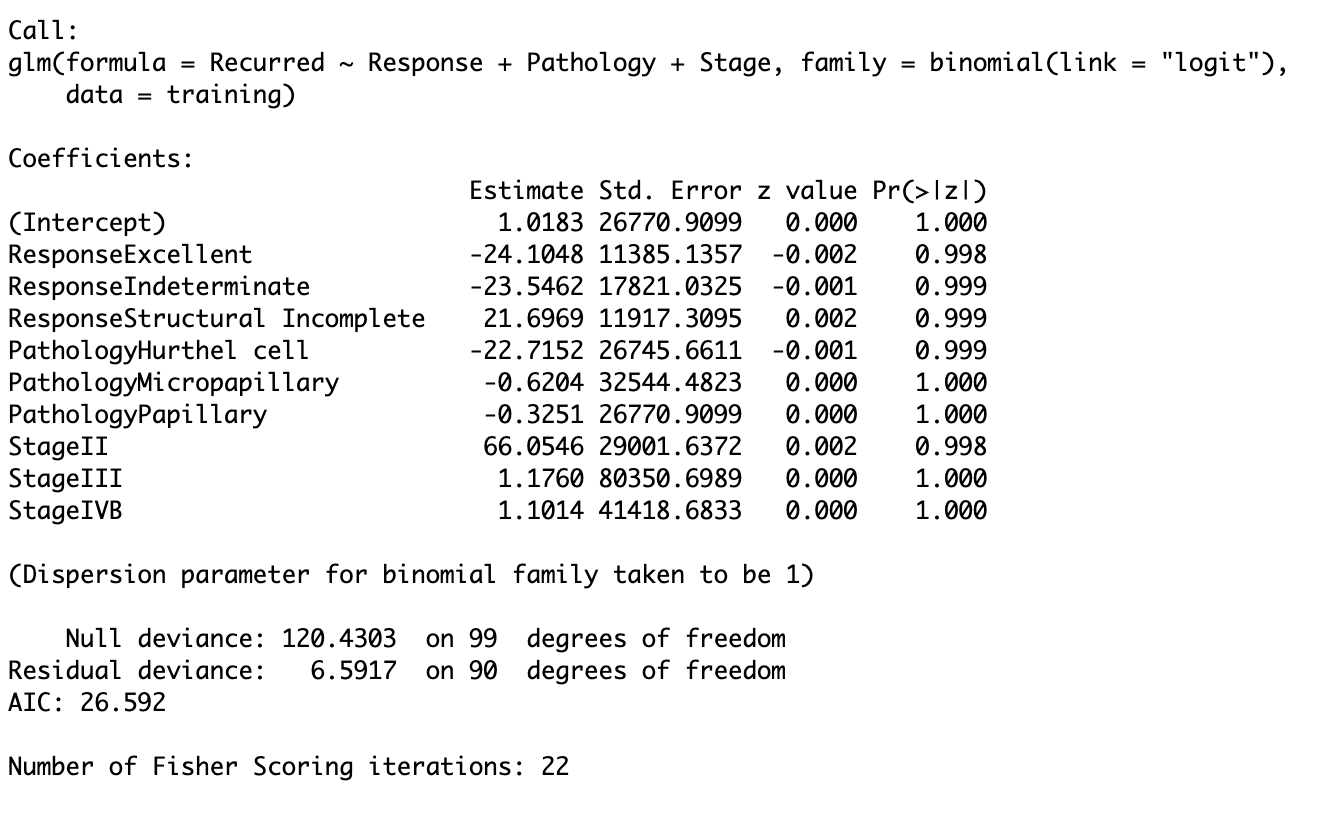


* + More involvement of lymph nodes may lead to a higher recurrence probability

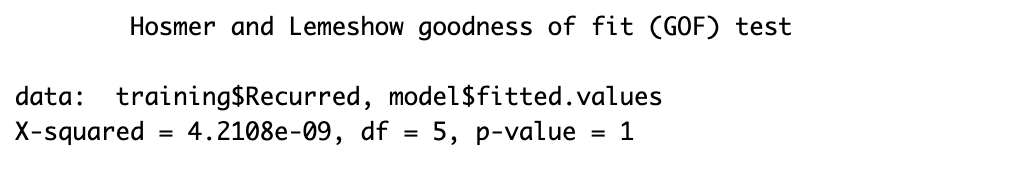


Remark: The percentage of recurrence is not completely reliable since some of the variables only have a small sample in particular categories. Therefore, we can fit the dataset into logistic regression and classification tree to see which of those potential explanatory variables are most responsible for the cancer recurrence.

1. Model and Classification
2. Logistic Regression



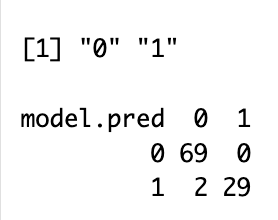
Logistic selects response, pathology, and stage as the three most important variables that influence whether the cancer will recur.



In *Hosmer and Lemeshow goodness of fit test,* the p-value is significantly large, showing that the model above is a good fit for the data.

Recurred ~ Response + Pathology + Stage

Using logistic regression as a classifier, we have an accuracy of 98% during testing.



1. Further explanation for “Response”, “Stage”, and “Pathology”