# P8130_final_project

Yuechu Hu, Leyang Rui, Yifei Yu, Jinghan Zhao

2024-12-03

## Abstract

## Introduction

The data we used for this analysis originates from a breast cancer survival dataset collected from a prospective study. The dataset contains 14 important variables, including patients' age, race, marital status, tumor size, cancer stages (T Stage, N Stage, and 6th Stage), estrogen and progesterone status, and regional node involvement. The dataset also records patients' survival times in months and their final survival status (dead or alive). Our objective is to develop models that predict the risk of death among breast cancer patients using these features. Specifically, we aim to identify which variables significantly impact patients' survival outcomes, determine (potential interactions among the variables, and assess the performance of the models across racial groups). Additionally, we will investigate fairness in the model predictions to ensure equitable accuracy between the majority race group "White" and the minority "Black".

## Methods

This report was conducted using `survival_data`, which contains information about breast cancer patients from a prospective study. The dimension of this dataset is 4024 rows and 16 columns. It contains five numeric variables, which are `age`, `tumor_size`, `regional_node_examined`, `regional_node_positive`, and `survival_months`, and eleven categorical variables, which are `race`, `marital_status`, `t_stage`, `n_stage`, `x6th_stage`, `differentiate`, `grade`, `a_stage`, `estrogen_status`, `progesterone_status`, and `status`. Key variables in the dataset include:

**age**: The age of the patient (in years).

**differentiate**: Tumor differentiation grade, categorized as "Well differentiated", "Moderately differentiated", "Poorly differentiated", or "Undifferentiated".

**a_stage**: Categorized as "Regional" (a neoplasm that has extended) or "Distant" (a neoplasm that has spread to parts of the body remote from).

**tumor_size**: The size of tumor (in millimeters).

**regional_node_examined**: The number of examined regional nodes.

**regional_node_positive**: The number of positive regional nodes.

**survival_month**: The time of a patient with breast cancer is expected to live after their diagnosis (in months).

The dataset comprises a total of 4024 observations, with no missing values in the primary variables analyzed. Firstly, after cleaning and tidying the data, we could figure out the distributions of the data and check for potential outliers or influential points by plotting the distributions of the variables, such as histograms and box plots. We also examine the pairwise relationships between variables.

## Results

As shown in figure 1, we can find out that most patients are between 40 and 70 years old, and most of the survival time are larger than 45 months. The number of examined regional nodes for most subjects are smaller than 30, and the subjects with nearly 12 examined regional nodes are the most. It is worth noting that the distributions of both variables `regional_node_positive` and `tumor_size` are significantly skewed to the right. Over 2500 subjects only have 1 or 2 positive regional nodes, which is the most frequent number of positive regional nodes. Most of the tumor sizes are smaller than 50 mm, and we can find that the most frequent size is around 19 mm, followed by around 14 mm.

The figure 2 distributed the survival time by the status, showing that survival months are significantly higher for the Alive group compared to the Dead group. According the figure 3, from T1 to T3, as the stage changes, both the mean tumor sizes and IQR become larger. At T4 stage, the IQR of tumor sizes is much larger than others, and the mean size is smaller than the mean size at T3 stage. We notice that there are some potential outliers both ar T1 stage and T3 stage. The survival time is longer in the Regional stage, and the Alive group shows higher survival times across both stages by looking at the figure 4. In the figure 5, the Undifferentiated group has larger tumor sizes compared to the other categories, while the Well, Moderately, and Poorly differentiated groups show similar distributions with many smaller tumors and numerous high-value outliers. The figure 6 highlights the differences in tumor size distribution and trends with age between individuals who are alive and those who are deceased. While the "Alive" group shows no significant relationship between age and tumor size, the "Dead" group exhibits a pattern where larger tumors are associated with younger ages.

According the figure 7, as it changes from undifferentiated to well differentiated, the negative correlation between the number of positive regional nodes and the survival months becomes weaker.

## Conclusion/Discussion

## Group members' contribution

Leyang and Jinghan focused on statistical methods, building models to extract meaningful insights. Yuechu and Yifei concentrated on data description and visualization, presenting information through clear visuals. All four of us contributed to the writing of the report, integrating our individual efforts into a cohesive final report that reflects both analytical depth and visual clarity.

## Appendix

Table 1: Summary Statistics for Numeric Variables

| Variable Name | Mean | SD | Median | IQR |
|---|---|---|---|---|
| Age | 53.972167 | 8.963134 | 54 | 14 |
| Tumor Size | 30.473658 | 21.119696 | 25 | 22 |
| Regional Nodes Examined | 14.357107 | 8.099675 | 14 | 10 |
| Regional Nodes Positive | 4.158052 | 5.109331 | 2 | 4 |
| Survival Months | 71.297962 | 22.921429 | 73 | 34 |

Table 2: Summary Statistics for Categorical Variables

| Variable Name | Level | Count | Proportion |
|---|---|---|---|
| Estrogen Status | Positive | 3755 | 0.9332 |
| Estrogen Status | Negative | 269 | 0.0668 |
| Progesterone Status | Positive | 3326 | 0.8265 |
| Progesterone Status | Negative | 698 | 0.1735 |
| Status | Alive | 3408 | 0.8469 |
| Status | Dead | 616 | 0.1531 |

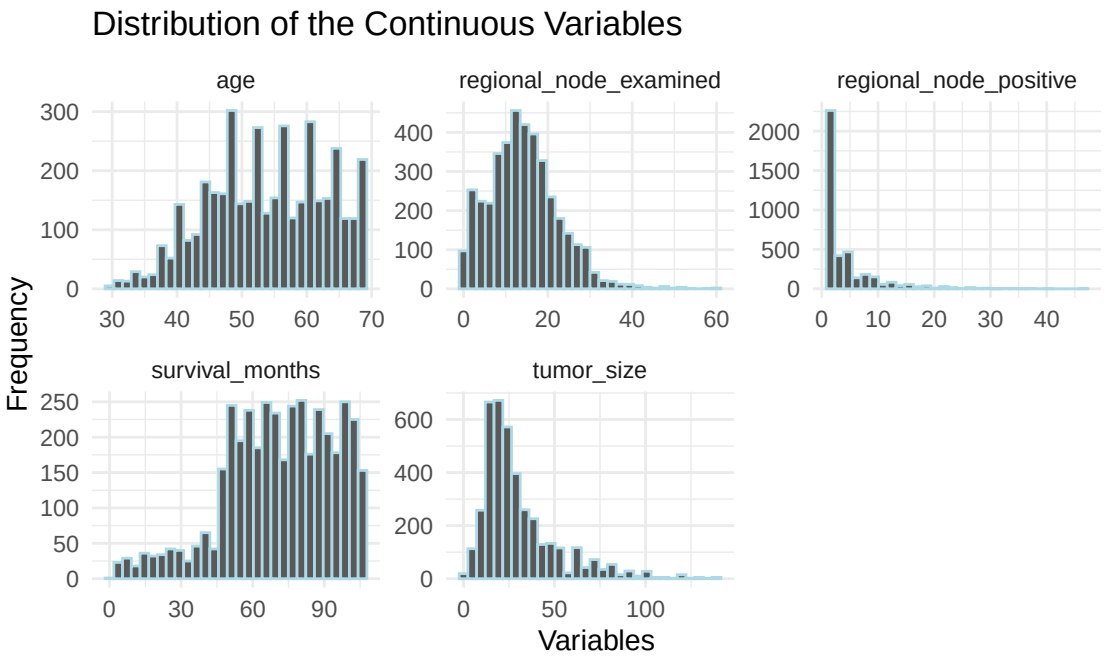


Figure 1: Distribution of the Continuous Variables

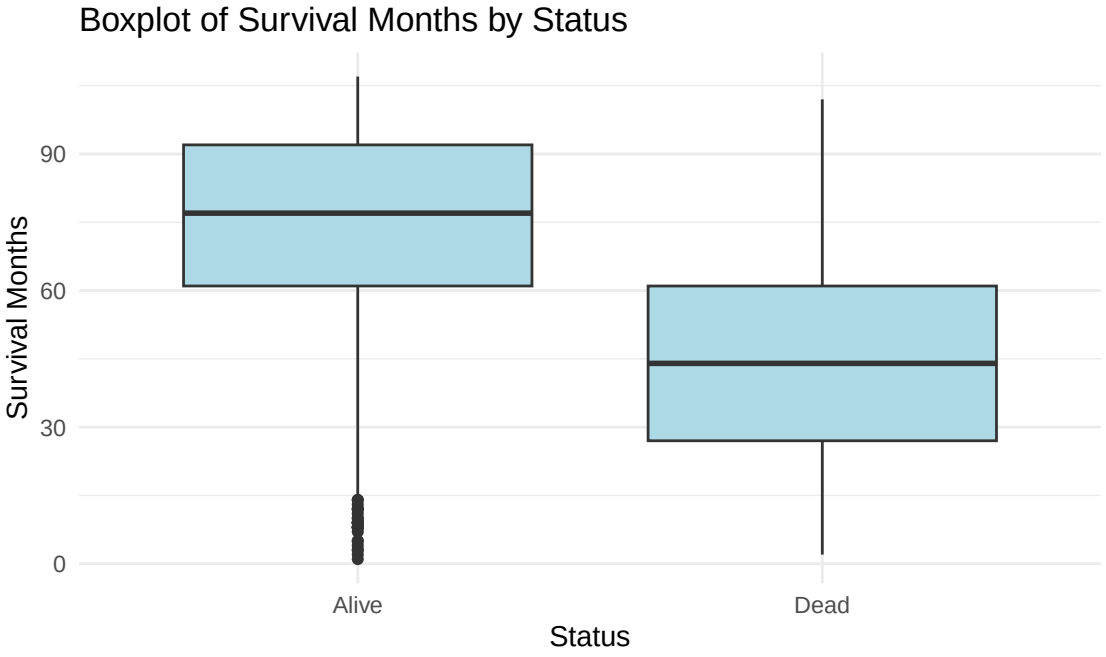


Figure 2: Survival Months by Status

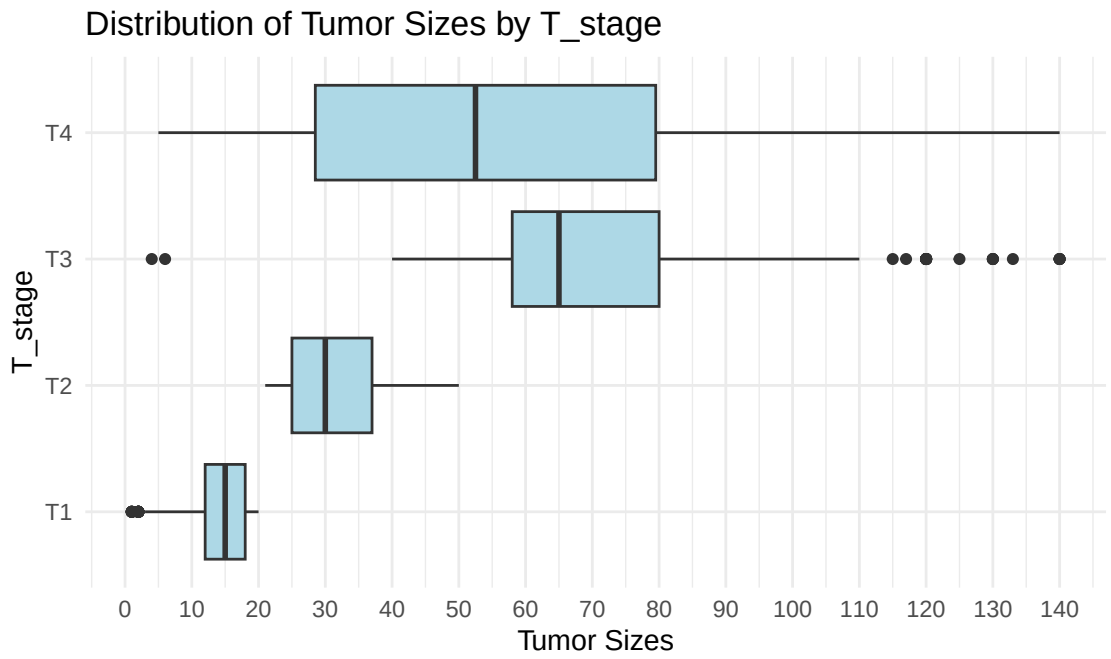


Figure 3: Tumor Sizes by T stage

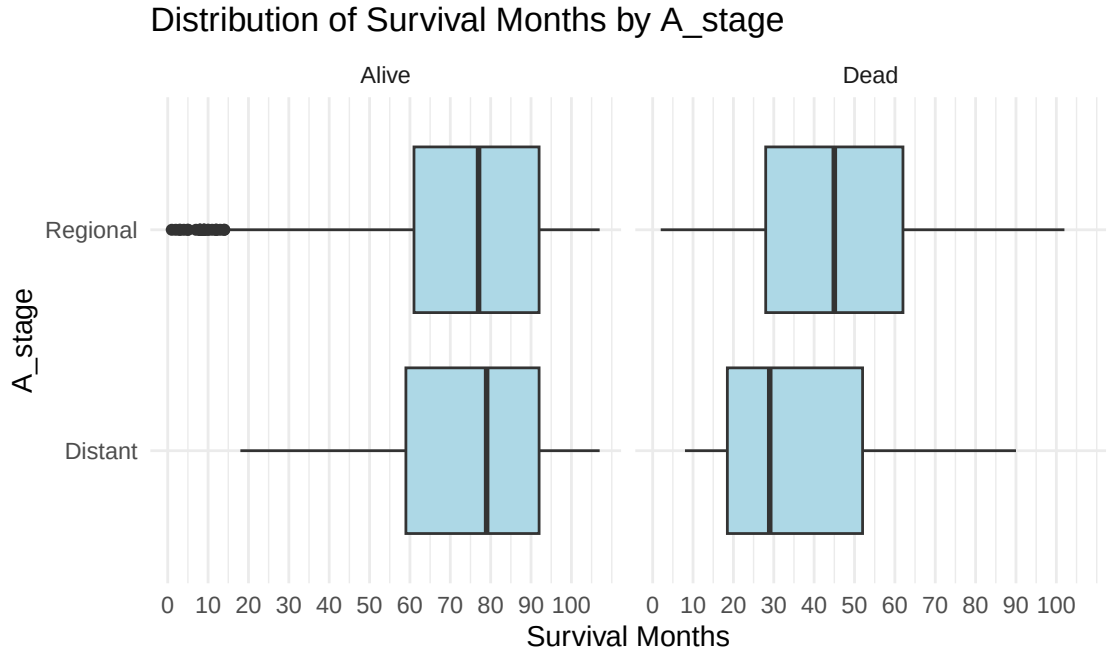


Figure 4: Survival Months by A stage Based on Status

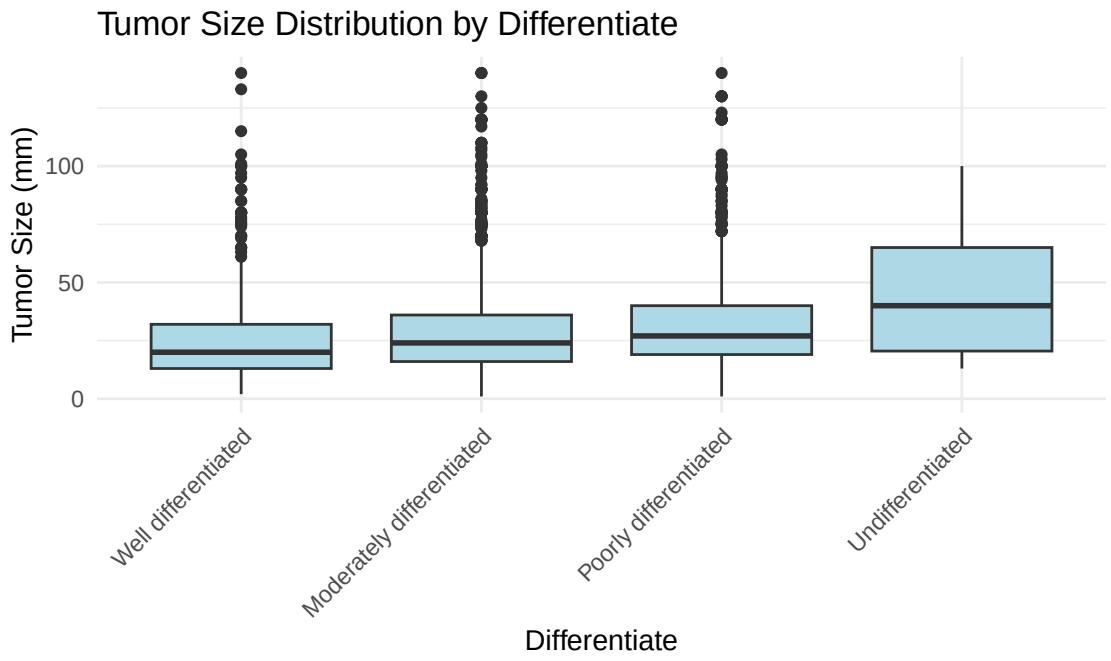


Figure 5: Tumor Size by Differentiate

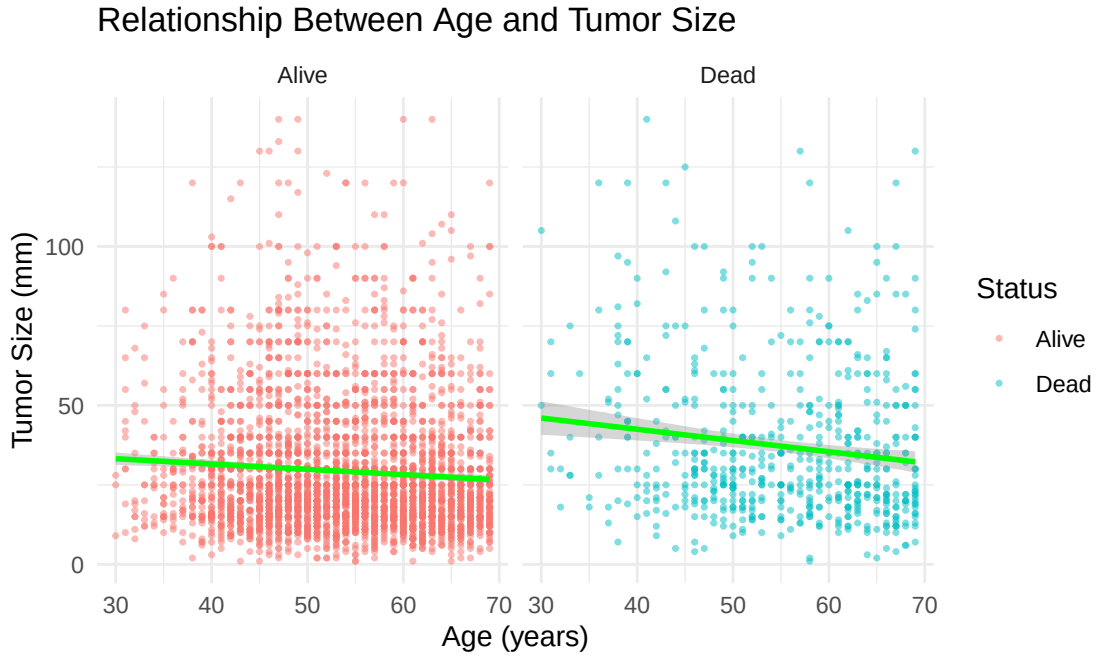


Figure 6: Relationship Between Age and Tumor Size across Status

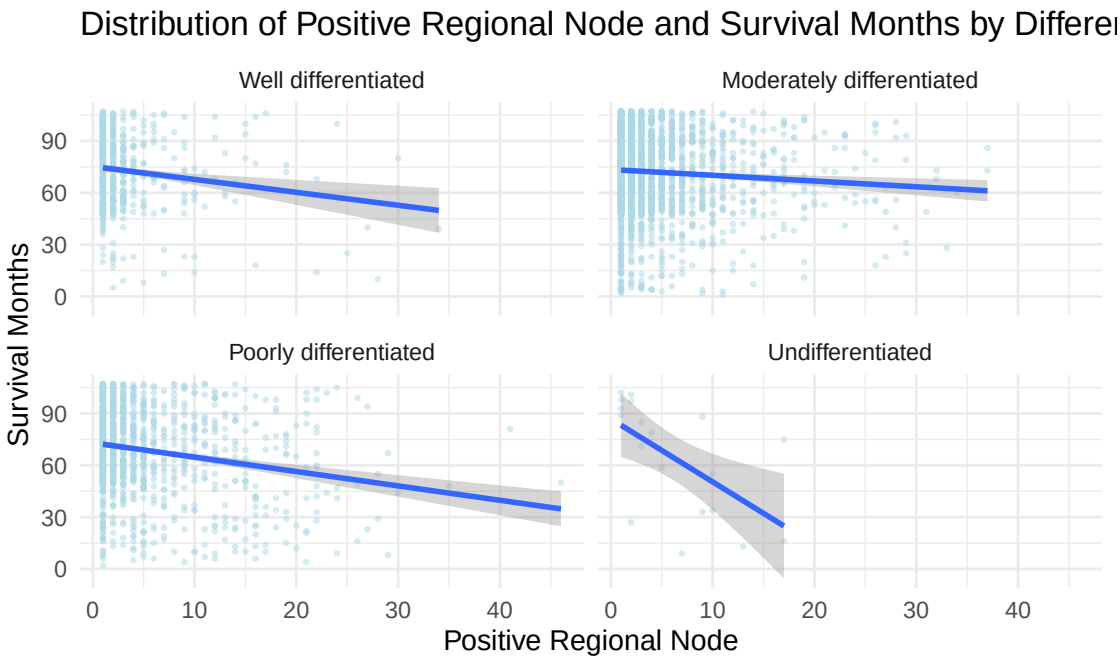


Figure 7: Positive Regional Node vs Survival Months Across Differentiate