biosta_final_raw

2024-12-20

Modeling Survival Outcomes in Breast Cancer Xiaoni Xu, Yiran Xu, Erin Ge, Boxiang Tang

Abstract Breast cancer is a leading cause of cancer-related mortality among women, making it essential to understand survival predictors and develop related prediction models. This study analyzed survival data from 4,024 breast cancer patients using demographic, clinical, and pathological variables. After data preprocessing, survival models, including log-logistic, Cox proportional hazards, Weibull, and lognormal models, were compared. The log-logistic model was identified as the best fit, with the lowest Akaike Information Criterion, outlining significant predictors such as age, tumor size, hormone receptor status, nodal involvement, and tumor stage. The findings suggest that the log-logistic model offers a suitable framework for understanding survival outcomes for breast cancer patients.

Introduction Breast cancer is one of the most prevalent forms of cancer globally and is among the leading causes of cancer-related mortality among women (Harbeck et al. 2019). Understanding the factors that influence survival outcomes and accurately predicting the risk are critical for improving patient care and clinical treatment. This report analyzes a dataset from a prospective study on breast cancer patients, aiming to explore predictors of survival and assess the fairness and performance of predictive models. The analysis begins with a comprehensive exploration of the dataset, including descriptive statistics and visualizations to understand the distribution and relationships between variables. We are able to find that the log-logistic model is superior compared to the Cox proportional hazards model. The report also addresses potential challenges, such as multicollinearity, model assumptions, and influential observations, ensuring robust conclusions.

Methods This study utilizes a dataset from a prospective study on breast cancer patients. The dataset includes detailed baseline information about demographic, clinical, and pathological features. Table 1 provides an overview of the variables included in the study, along with their descriptions and categories. The analysis began with comprehensive data cleaning and processing in R. Variables were renamed for consistency, categorical variables were converted to factors, and transformations were applied where appropriate to improve interpretability and address potential skewness. Exploratory data analysis (EDA) was conducted to understand the distributions and relationships within the dataset. Summary statistics were computed for all variables, and visualizations such as histograms and bar plots were created. Side-by-side histograms for categorical variables, such as Race, Marital Status, T Stage, and N Stage, showed their frequency distributions, and numerical variables like Age, Tumor Size, and Regional Nodes Examined were visualized using histograms to identify patterns or potential outliers (Figure 1). Figure 2 shows the distributions of log-transformed variables in order to stabilize variance. Outliers and influential points in the dataset were identified and removed to ensure the robustness of the regression model. Cook's Distance, a metric in regression analysis, was applied to detect influential observations. Cook's Distance evaluates both the leverage of a data point (its distance from the mean of the predictors) and its residual (the deviation of the observation from the fitted model). By combining these two measures, Cook's Distance highlights observations that disproportionately affect the model's estimates. To identify influential points, Cook's Distance was computed for each observation in the dataset using the fitted logistic regression model. Observations with a Cook's Distance exceeding the threshold of 4/n, where n is the total number of observations, were flagged as influential. A diagnostic plot was generated to visually inspect Cook's Distance values, with points above the threshold marked in red for clarity (Figure 3). Observations identified as influential were removed from the dataset before refitting the model. To assess multicollinearity among numerical variables, Variance Inflation Factor (VIF) was calculated. Multicollinearity occurs when two or more predictor variables are highly correlated, leading to unstable regression coefficients. VIF quantifies the extent of multicollinearity, with higher values indicating greater correlation. A VIF value above 5 is generally considered a sign of

significant multicollinearity. The numerical variables analyzed included age, log-transformed tumor size, log-transformed number of nodes examined, the number of regional nodes positive, and survival months. A linear regression model was fitted using these variables, and VIF values were calculated for each predictor. The results showed that all VIF values were below the threshold of 5 (Table 2). This indicates that no significant multicollinearity was present among the numerical variables.

Results For the original data, the analysis was conducted on a dataset containing 4,024 patients with 16 variables, representing demographic, clinical, and tumor-specific characteristics. The age of patients ranged from 30 to 69 years, with a median age of 54. Tumor size varied widely, from 1 mm to 140 mm, with an average size of 30.47 mm. The majority of patients (84.8%) identified as White, with Black and Other racial categories accounting for 7.2% and 8.0%, respectively. Most patients were married (65.7%), followed by being single (15.3%). Tumor staging variables (T stage and N stage) highlighted that 79% of patients had early-stage tumors classified as T1 or T2, while lymph node involvement (regional_node_positive) ranged from 1 to 46, with a mean of 4.16. The compatibility of the survival data with Weibull and lognormal parametric models was evaluated to determine the most appropriate distribution for the survival analysis. The Weibull model is a commonly used parametric survival model. The Anderson-Darling test was employed to assess the goodness-of-fit of the Weibull distribution to the data. The test produced an A_n statistic of 41.662 with a p-value of 1.636×10^{-7}, strongly rejecting the null hypothesis that the data follow a Weibull distribution. Similarly, a lognormal model was fitted to the survival data, and four plots were generated for both the lognormal and Weibull models to visually inspect the fit: Empirical and Theoretical Densities Plot, Q-Q Plot, Empirical and Theoretical CDFs plot, and P-P Plot (Figure 4). For both models, significant deviations from the diagonal in the Q-Q plot, suggesting poor fit and that the survival data are not well modeled by either the Weibull or the lognormal distribution. The log-logistic model, a parametric survival model that assumes survival times follow a log-logistic distribution, was fitted to the data using the survival months and status as the response variable. The model included all available predictors, and the fitting process yielded a log-likelihood of -954.3, significantly improved from the intercept-only model (-1295), with a chi-squared statistic of 681.29 (p-value $< 8\times10^{-125}$). This indicates that the model provides a better fit to the data compared to the null model. Predictors with $p > 0.05$ and non-significant effects were excluded, resulting in an updated model formula containing only significant variables. The model output is shown in Figure 5. The Cox proportional hazards model, a semi-parametric approach that does not assume a specific form for the baseline hazard, was fitted to the survival data. The global likelihood ratio test, Wald test, and score (logrank) test all yielded statistically significant results ($p < 2\times10^{-16}$), indicating that the model as a whole is predictive of survival. The concordance index (C=0.868) reflects strong predictive discrimination, suggesting that the model effectively distinguishes between high- and low-risk individuals. The model output is shown in Figure 6. The Akaike Information Criterion (AIC) was used to compare the relative quality of the log-logistic and Cox proportional hazards models. AIC evaluates the trade-off between goodness-of-fit and model complexity, with lower values indicating a better balance. The AIC for the log-logistic model was 1970.695, whereas the AIC for the Cox model was 4275.727. This significant difference suggests that the log-logistic model provides a more parsimonious fit to the data. The log-logistic model, therefore, is preferred based on the AIC criterion for modeling survival in this dataset.

Conclusions and Discussion This study analyzed survival data from breast cancer patients using various parametric and semi-parametric models, focusing on identifying key predictors and assessing the performance of survival models. The results provide valuable insights into the demographic, clinical, and pathological factors influencing survival outcomes, while also highlighting the relative strengths of different modeling approaches. The log-logistic model emerged as the most suitable for this dataset, as indicated by its lower Akaike Information Criterion (AIC) value compared to the Cox proportional hazards model. Key predictors identified, such as age, tumor size, hormone receptor status, nodal involvement, and tumor stage, align with established clinical understanding of breast cancer. The study also explored Weibull and lognormal models, which were found to be incompatible with the data based on goodness-of-fit tests and diagnostic plots. These findings underscore the importance of evaluating multiple modeling approaches and considering their assumptions and limitations in survival analysis. The findings of this study provide a foundation for improving risk prediction and personalized treatment strategies, emphasizing the essential role of statistical analysis in breast cancer research.

Group Member Contributions Xiaoni Xu rewrote the entire R code of the project; the preliminary models

were not used, and only histograms were shown for data visualization. Xiaoni Xu also generated all tables and figures of the report along with drafting and finishing all written sections of the report. Yiran Xu drafted the codes for the preliminary models, Weibull and lognormal models, and data cleaning; Erin Ge followed up with covariate further code editing including model output and outliers detection, as well as interpretation. Boxiang Tang built upon the preliminary modeling work (Logistic & Cox models) conducted by Xiaoni, adding a complete data manipulation and data exploration section (including a total 8 steps). Additionally, he enhanced the modeling sectionsetion (with a total of 7 steps), systematically organizing and completing the data analysis and modeling process. He also incorporated many numerous data exploration insights, model comparisonscomparisions, diagnostic correction, and calibration plots., along with detailed summaries of data and summary tables for the Logistics & COx models. edledited those aforementioned preliminary models with pie charts and boxplots.

References Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., & Cardoso, F. (2019). Breast cancer. Nature Reviews Disease Primers, 5, Article 66. https://doi.org/10.1038/s41572-019-0111-2

I. Data Manupulation & Exploration:

```
library(readr)
library(dplyr)
library(survival)
library(caret)
library(ggplot2)
library(patchwork)
library(kableExtra)
library(minerva)
library(stats)
library(fitdistrplus)
library(randomForestSRC)
library(tidyverse)
library(goftest)
library(survival)
library(stats)
library(broom)
library(car)
library(graphics)
library(MASS)
library(e1071)
library(pROC)
library(pec)
library(glmnet)
library(ROSE)
library(caret)
library(survminer)
library(kableExtra)
```

Step_1: Loading necessary packages

```r
data_clean = read_csv("/Users/boxiangtang/Desktop/Biosta_HW/Biosta_final/Project_2_data.csv") |>
  rename(
    Stage_6th = `6th Stage`) |>
  janitor::clean_names() |>

mutate(status = as.numeric(status == "Dead"),
       race = as.factor(race),
       marital_status = as.factor(marital_status),
       t_stage = as.factor(t_stage),
       n_stage = as.factor(n_stage),
       stage_6th = as.factor(stage_6th),
       differentiate = as.factor(differentiate),
       grade = as.factor(grade),
       a_stage = as.factor(a_stage),
       estrogen_status = as.factor(estrogen_status),
       progesterone_status = as.factor(progesterone_status),
       survival_months = log(survival_months) + 1
  )


# correct the col name: from "reginol_node_positive" to "regional_node_positive" (TBX)
names(data_clean)[names(data_clean) == "reginol_node_positive"] <- "regional_node_positive"


summary(data_clean)
```

Step_2: Data cleaning & Initial Survival Time Distribution Visualization

```
##       age           race        marital_status t_stage   n_stage   stage_6th
##  Min.   :30.00   Black: 291   Divorced : 486   T1:1603   N1:2732   IIA :1305
##  1st Qu.:47.00   Other: 320   Married  :2643   T2:1786   N2: 820   IIB :1130
##  Median :54.00   White:3413   Separated:  45   T3: 533   N3: 472   IIIA:1050
##  Mean   :53.97                Single   : 615   T4: 102             IIIB:  67
##  3rd Qu.:61.00                Widowed  : 235                       IIIC: 472
##  Max.   :69.00
##                    differentiate                    grade          a_stage
##  Moderately differentiated:2351   1                   : 543   Distant :  92
##  Poorly differentiated    :1111   2                   :2351   Regional:3932
##  Undifferentiated         :  19   3                   :1111
##  Well differentiated      : 543   anaplastic; Grade IV:  19
##
##
##    tumor_size     estrogen_status progesterone_status regional_node_examined
##  Min.   :  1.00   Negative: 269   Negative: 698       Min.   : 1.00
##  1st Qu.: 16.00   Positive:3755   Positive:3326       1st Qu.: 9.00
##  Median : 25.00                                       Median :14.00
##  Mean   : 30.47                                       Mean   :14.36
##  3rd Qu.: 38.00                                       3rd Qu.:19.00
##  Max.   :140.00                                       Max.   :61.00
##  regional_node_positive survival_months     status
##  Min.   : 1.000         Min.   :1.000   Min.   :0.0000
##  1st Qu.: 1.000         1st Qu.:5.025   1st Qu.:0.0000
##  Median : 2.000         Median :5.290   Median :0.0000
```

```
##  Mean   : 4.158         Mean   :5.184   Mean   :0.1531
##  3rd Qu.: 5.000         3rd Qu.:5.500   3rd Qu.:0.0000
##  Max.   :46.000         Max.   :5.673   Max.   :1.0000
```

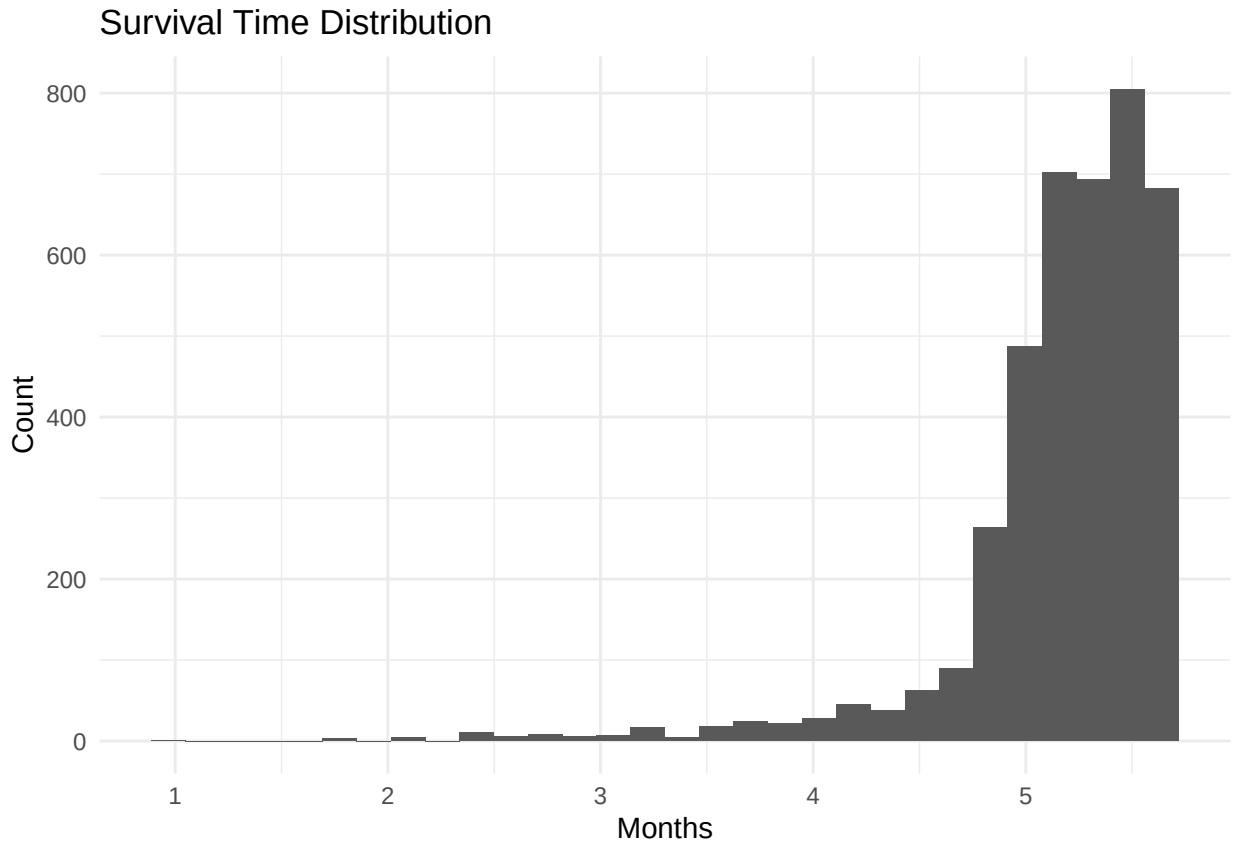
```
head(data_clean) |>
  kable() |>
  kable_styling(full_width = FALSE) %>%
  scroll_box(width = "100%", height = "300px")
```

age	race	marital_status	t_stage	n_stage	stage_6th	differentiate	grade	a_stage	tumor_s
68	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	
50	White	Married	T2	N2	IIIA	Moderately differentiated	2	Regional	
58	White	Divorced	T3	N3	IIIC	Moderately differentiated	2	Regional	
58	White	Married	T1	N1	IIA	Poorly differentiated	3	Regional	
47	White	Married	T2	N1	IIB	Poorly differentiated	3	Regional	
51	White	Single	T1	N1	IIA	Moderately differentiated	2	Regional	

```
summary_stats <- data_clean |> summary()
print(summary_stats)
```

```
##       age           race        marital_status t_stage   n_stage   stage_6th
##  Min.   :30.00   Black: 291   Divorced : 486   T1:1603   N1:2732   IIA :1305
##  1st Qu.:47.00   Other: 320   Married  :2643   T2:1786   N2: 820   IIB :1130
##  Median :54.00   White:3413   Separated:  45   T3: 533   N3: 472   IIIA:1050
##  Mean   :53.97                Single   : 615   T4: 102             IIIB:  67
##  3rd Qu.:61.00                Widowed  : 235                       IIIC: 472
##  Max.   :69.00
##                    differentiate                    grade          a_stage
##  Moderately differentiated:2351   1                   : 543   Distant :  92
##  Poorly differentiated    :1111   2                   :2351   Regional:3932
##  Undifferentiated         :  19   3                   :1111
##  Well differentiated      : 543   anaplastic; Grade IV:  19
##
##
##    tumor_size     estrogen_status progesterone_status regional_node_examined
##  Min.   :  1.00   Negative: 269   Negative: 698       Min.   : 1.00
##  1st Qu.: 16.00   Positive:3755   Positive:3326       1st Qu.: 9.00
##  Median : 25.00                                       Median :14.00
##  Mean   : 30.47                                       Mean   :14.36
##  3rd Qu.: 38.00                                       3rd Qu.:19.00
##  Max.   :140.00                                       Max.   :61.00
##  regional_node_positive survival_months     status
##  Min.   : 1.000         Min.   :1.000   Min.   :0.0000
##  1st Qu.: 1.000         1st Qu.:5.025   1st Qu.:0.0000
##  Median : 2.000         Median :5.290   Median :0.0000
##  Mean   : 4.158         Mean   :5.184   Mean   :0.1531
##  3rd Qu.: 5.000         3rd Qu.:5.500   3rd Qu.:0.0000
##  Max.   :46.000         Max.   :5.673   Max.   :1.0000
```

```r
data_clean |> ggplot(aes(x = survival_months, fill = status)) +
  geom_histogram() +
  labs(title = "Survival Time Distribution", x = "Months", y = "Count") +
  theme_minimal()
```




```r
# Create a data frame for variable descriptions
variable_descriptions <- data.frame(
  Variable = c(
    "Age", "Race", "Marital Status", "T Stage (Tumor)", "N Stage (Node)",
    "Stage (6th Edition)", "Differentiate", "Grade", "A Stage",
    "Tumor Size", "Estrogen Status", "Progesterone Status",
    "Regional Nodes Examined", "Regional Nodes Positive",
    "Survival Months", "Status"
  ),
  Description = c(
    "Patient's age at the time of diagnosis or study enrollment.",
    "Patient's racial identity: Black, White, Other.",
    "Patient's marital status: Married, Single, Divorced, Separated, Widowed.",
    "Tumor size and extent: T1 ( 2 cm), T2 (>2 cm but  5 cm), T3 (>5 cm), T4 (invasion into chest wall o
    "Lymph node involvement: N1 (1-3 nodes), N2 (4-9 nodes), N3 ( 10 nodes).",
    "Overall cancer stage: IIA, IIB, IIIA, IIIB, IIIC.",
    "Tumor differentiation level: Well, Moderately, Poorly, Undifferentiated.",
```

```
    "Tumor histological grade: Grade 1 (low), Grade 2 (moderate), Grade 3 (high), Grade IV (anaplastic).",
    "Extent of cancer spread: Regional (local spread), Distant (metastasized).",
    "Size of the tumor in millimeters.",
    "Tumor's estrogen receptor status: Positive, Negative.",
    "Tumor's progesterone receptor status: Positive, Negative.",
    "Number of regional lymph nodes examined for cancer.",
    "Number of regional lymph nodes found to be cancer-positive.",
    "Number of months the patient survived after diagnosis or study enrollment.",
    "Patient's status at the end of the study: Alive, Deceased."
  )
)

# Display the table
kable(variable_descriptions, col.names = c("Variable", "Description"),
      caption = "Variable Descriptions in the Breast Cancer Dataset", align = "l")
```

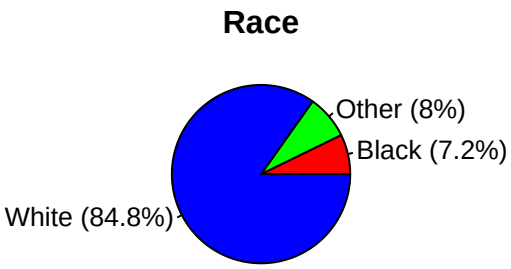
Step_3: Variable Description Tables

Table 2: Variable Descriptions in the Breast Cancer Dataset

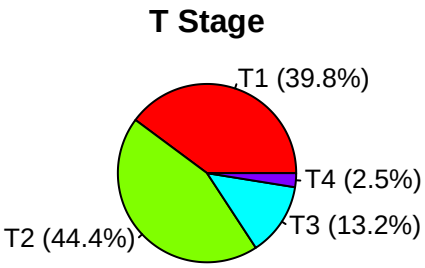
Variable	Description
Age	Patient's age at the time of diagnosis or study enrollment.
Race	Patient's racial identity: Black, White, Other.
Marital Status	Patient's marital status: Married, Single, Divorced, Separated, Widowed.
T Stage (Tumor)	Tumor size and extent: T1 (2 cm), T2 (>2 cm but 5 cm), T3 (>5 cm), T4 (invasion into chest wall or skin).
N Stage (Node)	Lymph node involvement: N1 (1–3 nodes), N2 (4–9 nodes), N3 (10 nodes).
Stage (6th Edition)	Overall cancer stage: IIA, IIB, IIIA, IIIB, IIIC.
Differentiate	Tumor differentiation level: Well, Moderately, Poorly, Undifferentiated.
Grade	Tumor histological grade: Grade 1 (low), Grade 2 (moderate), Grade 3 (high), Grade IV (anaplastic).
A Stage	Extent of cancer spread: Regional (local spread), Distant (metastasized).
Tumor Size	Size of the tumor in millimeters.
Estrogen Status	Tumor's estrogen receptor status: Positive, Negative.
Progesterone Status	Tumor's progesterone receptor status: Positive, Negative.
Regional Nodes Examined	Number of regional lymph nodes examined for cancer.
Regional Nodes Positive	Number of regional lymph nodes found to be cancer-positive.
Survival Months	Number of months the patient survived after diagnosis or study enrollment.
Status	Patient's status at the end of the study: Alive, Deceased.

```
# Ensure variables are converted to factor type
convert_to_factor <- function(data, vars) {
  data %>% mutate(across(all_of(vars), as.factor))
}
```

```r
# Convert variables to factor type
factor_vars <- c("race", "marital_status", "t_stage", "n_stage",
                 "stage_6th", "differentiate", "grade", "a_stage",
                 "estrogen_status", "progesterone_status", "status")
data_clean <- convert_to_factor(data_clean, factor_vars)

# Define a function to simplify long labels
simplify_labels <- function(label, max_length = 12) {
  ifelse(nchar(label) > max_length, paste0(substr(label, 1, max_length), "..."), label)
}

# Define a function to plot pie charts
plot_pie <- function(data, var, title) {
  counts <- data %>% count(!!sym(var)) %>% mutate(percent = round(n / sum(n) * 100, 1))
  counts <- counts %>% mutate(label = paste0(simplify_labels(as.character(!!sym(var))), " (", percent, "
  pie_data <- counts$n
  labels <- counts$label
  pie(pie_data, labels = labels, main = title, col = rainbow(length(pie_data)))
}

# Display in multiple pages
# Page 1
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))
plot_pie(data_clean, "race", "Race")
plot_pie(data_clean, "marital_status", "Marital Status")
plot_pie(data_clean, "t_stage", "T Stage")
plot_pie(data_clean, "n_stage", "N Stage")
```

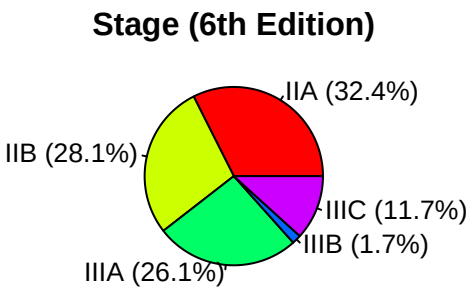


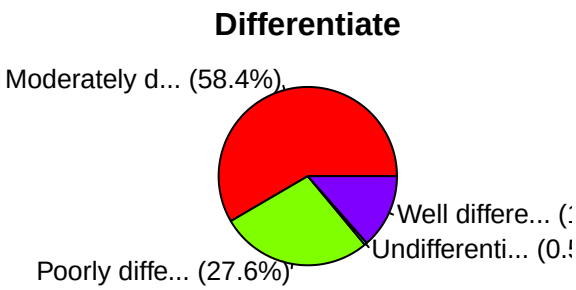
Married

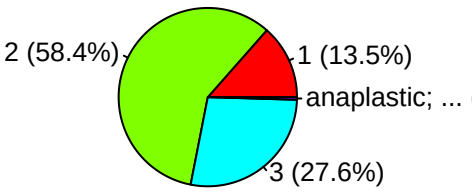


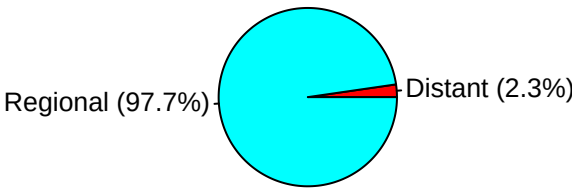
Step_4: Visualization for Categorical Variables

```
# Page 2
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))
plot_pie(data_clean, "stage_6th", "Stage (6th Edition)")
plot_pie(data_clean, "differentiate", "Differentiate")
plot_pie(data_clean, "grade", "Grade")
plot_pie(data_clean, "a_stage", "A Stage")
```

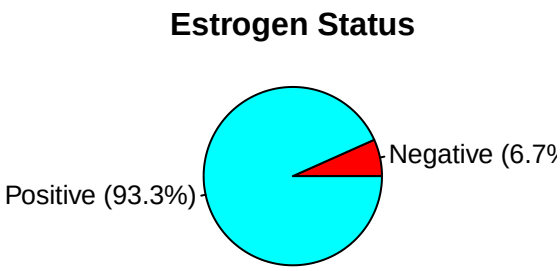


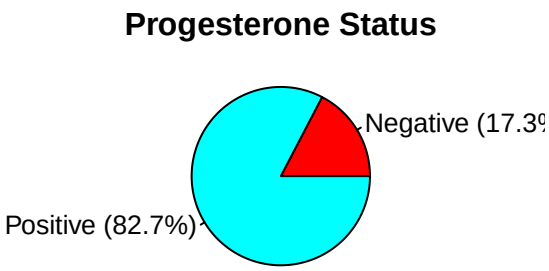


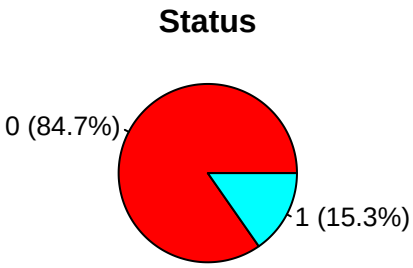





```
# Page 3
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))
plot_pie(data_clean, "estrogen_status", "Estrogen Status")
plot_pie(data_clean, "progesterone_status", "Progesterone Status")
plot_pie(data_clean, "status", "Status")
```

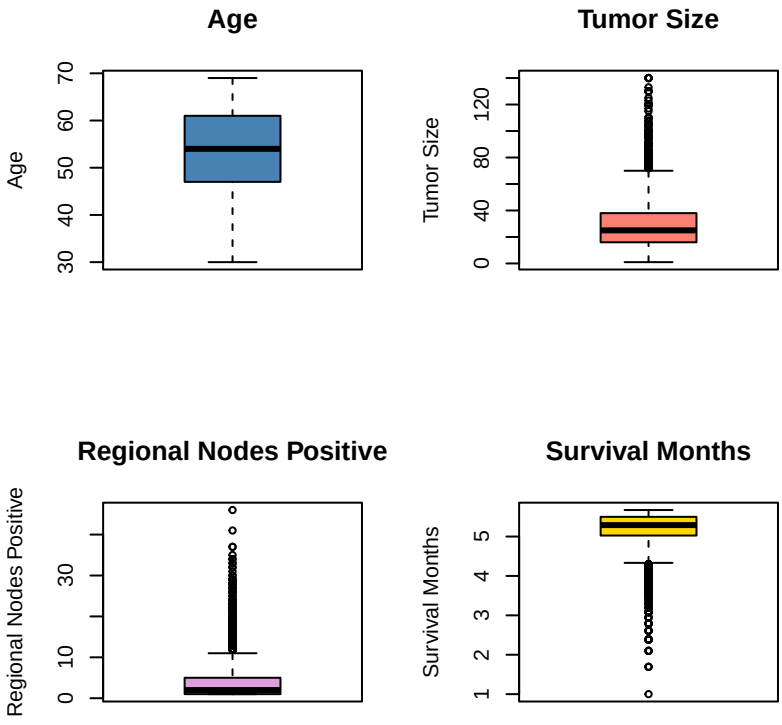







```r
# Select numeric variables and remove missing values
num_vars <- c("age", "tumor_size", "regional_node_examined", "regional_node_positive", "survival_months

# Use base R to filter the required columns
data_clean_numeric <- data_clean[, num_vars, drop = FALSE]

# Remove rows with missing values
data_clean_numeric <- na.omit(data_clean_numeric)

# Plot boxplots for each numeric variable
par(mfrow = c(2, 3)) # Set layout to 2 rows and 3 columns

boxplot(data_clean_numeric$age,
        main = "Age",
        ylab = "Age",
        col = "steelblue",
        border = "black")

boxplot(data_clean_numeric$tumor_size,
        main = "Tumor Size",
        ylab = "Tumor Size",
        col = "salmon",
        border = "black")
```

```r
boxplot(data_clean_numeric$regional_node_examined,
        main = "Regional Nodes Examined",
        ylab = "Regional Nodes Examined",
        col = "lightgreen",
        border = "black")

boxplot(data_clean_numeric$regional_node_positive,
        main = "Regional Nodes Positive",
        ylab = "Regional Nodes Positive",
        col = "plum",
        border = "black")

boxplot(data_clean_numeric$survival_months,
        main = "Survival Months",
        ylab = "Survival Months",
        col = "gold",
        border = "black")

# Reset plot parameters to default
par(mfrow = c(1, 1))
```



Step_5: Visualization for Numerical Variables

Step_6: Check Multi-collinearity between Different Kinds of Variable Checking Multi-Collinearity Among Numerical Variables(Correlation Matrix & Variance Inflation Factor (VIF)) :

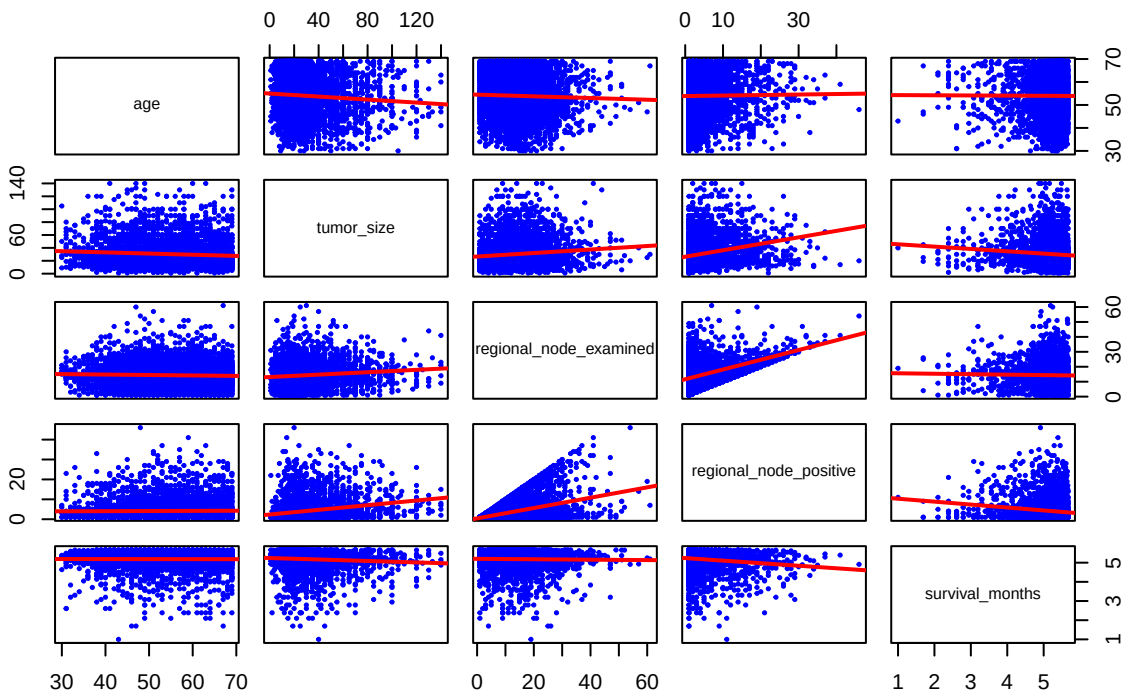
```r
# Select numerical variables
numeric_vars <- data_clean[, c("age", "tumor_size", "regional_node_examined", "regional_node_positive",

# Correlation matrix
cor_matrix <- cor(numeric_vars, use = "complete.obs")
print(cor_matrix)
```

```
##                              age  tumor_size regional_node_examined
## age                   1.000000000 -0.07721497            -0.03334548
## tumor_size           -0.077214971  1.00000000             0.10435180
## regional_node_examined -0.033345483  0.10435180             1.00000000
## regional_node_positive  0.012585513  0.24232172             0.41157970
## survival_months       -0.004077672 -0.08317533            -0.01816078
##                       regional_node_positive survival_months
## age                               0.01258551    -0.004077672
## tumor_size                        0.24232172    -0.083175332
## regional_node_examined            0.41157970    -0.018160776
## regional_node_positive            1.00000000    -0.139962706
## survival_months                  -0.13996271     1.000000000
```

```r
# Compute VIF
numeric_model <- lm(survival_months ~ ., data = numeric_vars)
vif_values <- vif(numeric_model)
print(vif_values)
```

```
##                    age             tumor_size regional_node_examined
##               1.008831               1.069709               1.206105
## regional_node_positive
##               1.267895
```

```r
# Select numerical variables (adjust to your dataset)
numerical_vars <- data_clean[, c("age","tumor_size", "regional_node_examined", "regional_node_positive"

# Create a scatterplot matrix
pairs(
  numerical_vars,
  panel = function(x, y) {
    points(x, y, pch = 20, col = "blue", cex = 0.5)  # Add scatterplot points
    abline(lm(y ~ x), col = "red", lwd = 2)          # Add red trend line
  },
  main = "Scatterplot Matrix with Trend Lines"
)
```



1. Numerical Variables - Covariance Matrix: Weak correlations between variables, with the highest at ~0.41 (regional_node_examined and regional_node_positive). - VIF Results: All values <2, indicating no multicollinearity. - Decision: Retain all numerical variables for now, as there are no strong correlations or multicollinearity issues.

Checking Multi-Collinearity Among Categorical Variables (Chi-Square Test of Independence):

```
# Select categorical variables
categorical_vars <- data_clean[, c("race",
                                   "marital_status",
                                   "t_stage",
                                   "n_stage",
                                   "stage_6th",
                                   "differentiate",
                                   "estrogen_status",
                                   "grade",
                                   "a_stage",
                                   "progesterone_status",
                                   "status")]

# Chi-Square or Fisher's Exact Test function with enhancements
perform_chi_square <- function(data, vars) {
  results <- data.frame(Variable1 = character(),
                        Variable2 = character(),
                        Test_Type = character(),
                        Statistic = numeric(),
                        P_Value = numeric())
```

```r
  for (i in 1:(length(vars) - 1)) {
    for (j in (i + 1):length(vars)) {
      var1 <- vars[i]
      var2 <- vars[j]
      table <- table(data[[var1]], data[[var2]])

      # Check if expected counts are too low for Chi-Square test
      if (any(chisq.test(table, simulate.p.value = TRUE)$expected < 5)) {
        # Use Fisher's Exact Test with Monte Carlo simulation if needed
        test <- fisher.test(table, simulate.p.value = TRUE, B = 10000)
        test_type <- "Fisher's Exact Test (Monte Carlo)"
        statistic <- NA  # Fisher's test does not produce a Chi-Square statistic
      } else {
        # Use Chi-Square Test
        test <- suppressWarnings(chisq.test(table))  # Suppress warnings for small expected counts
        test_type <- "Chi-Square Test"
        statistic <- test$statistic
      }

      # Append results
      results <- rbind(results, data.frame(Variable1 = var1,
                                           Variable2 = var2,
                                           Test_Type = test_type,
                                           Statistic = statistic,
                                           P_Value = test$p.value))
    }
  }

  return(results)
}

# Run the modified function
chi_square_results <- perform_chi_square(data_clean, colnames(categorical_vars))

# Display all results
print(chi_square_results)
```

```
##                       Variable1           Variable2
## 1                          race      marital_status
## X-squared                  race             t_stage
## X-squared1                 race             n_stage
## 11                         race           stage_6th
## 12                         race       differentiate
## X-squared2                 race     estrogen_status
## 13                         race               grade
## X-squared3                 race             a_stage
## X-squared4                 race progesterone_status
## X-squared5                 race              status
## 14             marital_status             t_stage
## X-squared6     marital_status             n_stage
## 15             marital_status           stage_6th
## 16             marital_status       differentiate
## 17             marital_status     estrogen_status
```

```
## 18              marital_status               grade
## 19              marital_status             a_stage
## X-squared7      marital_status progesterone_status
## X-squared8      marital_status              status
## X-squared9             t_stage             n_stage
## 110                    t_stage           stage_6th
## 111                    t_stage       differentiate
## X-squared10            t_stage     estrogen_status
## 112                    t_stage               grade
## 113                    t_stage             a_stage
## X-squared11            t_stage progesterone_status
## X-squared12            t_stage              status
## X-squared13            n_stage           stage_6th
## 114                    n_stage       differentiate
## X-squared14            n_stage     estrogen_status
## 115                    n_stage               grade
## X-squared15            n_stage             a_stage
## X-squared16            n_stage progesterone_status
## X-squared17            n_stage              status
## 116                stage_6th       differentiate
## 117                stage_6th     estrogen_status
## 118                stage_6th               grade
## 119                stage_6th             a_stage
## X-squared18        stage_6th progesterone_status
## X-squared19        stage_6th              status
## 120             differentiate     estrogen_status
## 121             differentiate               grade
## 122             differentiate             a_stage
## 123             differentiate progesterone_status
## 124             differentiate              status
## 125           estrogen_status               grade
## X-squared20   estrogen_status             a_stage
## X-squared21   estrogen_status progesterone_status
## X-squared22   estrogen_status              status
## 126                     grade             a_stage
## 127                     grade progesterone_status
## 128                     grade              status
## X-squared23           a_stage progesterone_status
## X-squared24           a_stage              status
## X-squared25 progesterone_status              status
##                                 Test_Type    Statistic       P_Value
## 1           Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## X-squared                    Chi-Square Test    8.4624312  2.061430e-01
## X-squared1                   Chi-Square Test    6.0796839  1.932759e-01
## 11          Fisher's Exact Test (Monte Carlo)           NA  3.513649e-01
## 12          Fisher's Exact Test (Monte Carlo)           NA  1.999800e-04
## X-squared2                   Chi-Square Test   13.4089972  1.225387e-03
## 13          Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## X-squared3                   Chi-Square Test    0.3069776  8.577104e-01
## X-squared4                   Chi-Square Test    5.0431477  8.033308e-02
## X-squared5                   Chi-Square Test   27.9700066  8.440929e-07
## 14          Fisher's Exact Test (Monte Carlo)           NA  1.232877e-01
## X-squared6                   Chi-Square Test   22.3525223  4.303021e-03
## 15          Fisher's Exact Test (Monte Carlo)           NA  4.049595e-02
```

```
## 16          Fisher's Exact Test (Monte Carlo)           NA  6.749325e-02
## 17          Fisher's Exact Test (Monte Carlo)           NA  1.348865e-01
## 18          Fisher's Exact Test (Monte Carlo)           NA  6.589341e-02
## 19          Fisher's Exact Test (Monte Carlo)           NA  1.041896e-01
## X-squared7                   Chi-Square Test   11.0468957  2.604200e-02
## X-squared8                   Chi-Square Test   28.2638125  1.102769e-05
## X-squared9                   Chi-Square Test  323.4137132  7.823527e-67
## 110         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 111         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## X-squared10                  Chi-Square Test   19.5498593  2.103929e-04
## 112         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 113         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## X-squared11                  Chi-Square Test   13.8082347  3.178147e-03
## X-squared12                  Chi-Square Test  103.4763086  2.779095e-22
## X-squared13                  Chi-Square Test 6686.8340572  0.000000e+00
## 114         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## X-squared14                  Chi-Square Test   42.5230811  5.837545e-10
## 115         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## X-squared15                  Chi-Square Test  355.5764205  6.131424e-78
## X-squared16                  Chi-Square Test   36.8460037  9.976816e-09
## X-squared17                  Chi-Square Test  269.9291427  2.430141e-59
## 116         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 117         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 118         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 119         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## X-squared18                  Chi-Square Test   42.4854392  1.323098e-08
## X-squared19                  Chi-Square Test  281.6484425  9.830332e-60
## 120         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 121         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 122         Fisher's Exact Test (Monte Carlo)           NA  2.429757e-02
## 123         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 124         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 125         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## X-squared20                  Chi-Square Test   15.5892186  7.870207e-05
## X-squared21                  Chi-Square Test 1054.8431237 2.156063e-231
## X-squared22                  Chi-Square Test  135.1557391  3.052608e-31
## 126         Fisher's Exact Test (Monte Carlo)           NA  2.079792e-02
## 127         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## 128         Fisher's Exact Test (Monte Carlo)           NA  9.999000e-05
## X-squared23                  Chi-Square Test    2.3828061  1.226770e-01
## X-squared24                  Chi-Square Test   35.7647266  2.226426e-09
## X-squared25                  Chi-Square Test  124.8853879  5.392080e-29
```

```r
# Filter significant results
significant_results <- chi_square_results %>%
  filter(P_Value < 0.05)

# Print significant results
print("Significant Associations:")
```

```
## [1] "Significant Associations:"
```

```r
print(significant_results)
```

```
##                       Variable1           Variable2
## 1                          race      marital_status
## 12                         race       differentiate
## X-squared2                 race     estrogen_status
## 13                         race               grade
## X-squared5                 race              status
## X-squared6       marital_status             n_stage
## 15               marital_status           stage_6th
## X-squared7       marital_status progesterone_status
## X-squared8       marital_status              status
## X-squared9              t_stage             n_stage
## 110                     t_stage           stage_6th
## 111                     t_stage       differentiate
## X-squared10             t_stage     estrogen_status
## 112                     t_stage               grade
## 113                     t_stage             a_stage
## X-squared11             t_stage progesterone_status
## X-squared12             t_stage              status
## X-squared13             n_stage           stage_6th
## 114                     n_stage       differentiate
## X-squared14             n_stage     estrogen_status
## 115                     n_stage               grade
## X-squared15             n_stage             a_stage
## X-squared16             n_stage progesterone_status
## X-squared17             n_stage              status
## 116                   stage_6th       differentiate
## 117                   stage_6th     estrogen_status
## 118                   stage_6th               grade
## 119                   stage_6th             a_stage
## X-squared18           stage_6th progesterone_status
## X-squared19           stage_6th              status
## 120               differentiate     estrogen_status
## 121               differentiate               grade
## 122               differentiate             a_stage
## 123               differentiate progesterone_status
## 124               differentiate              status
## 125             estrogen_status               grade
## X-squared20     estrogen_status             a_stage
## X-squared21     estrogen_status progesterone_status
## X-squared22     estrogen_status              status
## 126                       grade             a_stage
## 127                       grade progesterone_status
## 128                       grade              status
## X-squared24             a_stage              status
## X-squared25 progesterone_status              status
##                                  Test_Type  Statistic       P_Value
## 1           Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 12          Fisher's Exact Test (Monte Carlo)         NA  1.999800e-04
## X-squared2                    Chi-Square Test   13.40900  1.225387e-03
## 13          Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## X-squared5                    Chi-Square Test   27.97001  8.440929e-07
```

```
## X-squared6                   Chi-Square Test   22.35252  4.303021e-03
## 15          Fisher's Exact Test (Monte Carlo)         NA  4.049595e-02
## X-squared7                   Chi-Square Test   11.04690  2.604200e-02
## X-squared8                   Chi-Square Test   28.26381  1.102769e-05
## X-squared9                   Chi-Square Test  323.41371  7.823527e-67
## 110         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 111         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## X-squared10                  Chi-Square Test   19.54986  2.103929e-04
## 112         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 113         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## X-squared11                  Chi-Square Test   13.80823  3.178147e-03
## X-squared12                  Chi-Square Test  103.47631  2.779095e-22
## X-squared13                  Chi-Square Test 6686.83406  0.000000e+00
## 114         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## X-squared14                  Chi-Square Test   42.52308  5.837545e-10
## 115         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## X-squared15                  Chi-Square Test  355.57642  6.131424e-78
## X-squared16                  Chi-Square Test   36.84600  9.976816e-09
## X-squared17                  Chi-Square Test  269.92914  2.430141e-59
## 116         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 117         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 118         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 119         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## X-squared18                  Chi-Square Test   42.48544  1.323098e-08
## X-squared19                  Chi-Square Test  281.64844  9.830332e-60
## 120         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 121         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 122         Fisher's Exact Test (Monte Carlo)         NA  2.429757e-02
## 123         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 124         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 125         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## X-squared20                  Chi-Square Test   15.58922  7.870207e-05
## X-squared21                  Chi-Square Test 1054.84312 2.156063e-231
## X-squared22                  Chi-Square Test  135.15574  3.052608e-31
## 126         Fisher's Exact Test (Monte Carlo)         NA  2.079792e-02
## 127         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## 128         Fisher's Exact Test (Monte Carlo)         NA  9.999000e-05
## X-squared24                  Chi-Square Test   35.76473  2.226426e-09
## X-squared25                  Chi-Square Test  124.88539  5.392080e-29
```

2. Categorical Variables - Significant Relationships: Key pairs like race and marital_status, t_stage and n_stage, and progesterone_status and status show strong associations (p-value < 0.05). - Decision: Retain all significant categorical variables. Consider redundancy (e.g., between t_stage and n_stage) in further analysis.

Checking Multi-Collinearity Between Numerical and Categorical Variables (ANOVA):

```r
perform_anova <- function(data, numeric_vars, categorical_vars) {
  results <- data.frame(Numeric_Var = character(), Categorical_Var = character(), P_Value = numeric())
  for (num_var in numeric_vars) {
    for (cat_var in categorical_vars) {
      formula <- as.formula(paste(num_var, "~", cat_var))
      anova_result <- anova(lm(formula, data = data))
      results <- rbind(results, data.frame(Numeric_Var = num_var, Categorical_Var = cat_var, P_Value = a
```

```
    }
  }
  results
}

numeric_vars <- c("age", "tumor_size", "regional_node_examined", "regional_node_positive", "survival_mo

categorical_vars <- c("race",
                      "marital_status",
                      "t_stage",
                      "n_stage",
                      'stage_6th',
                      "differentiate",
                      "estrogen_status",
                      "grade",
                      "a_stage",
                      "progesterone_status",
                      "status")

anova_results <- perform_anova(data_clean, numeric_vars, categorical_vars)
print(anova_results)
```

```
##               Numeric_Var     Categorical_Var       P_Value
## 1                     age                race  4.816383e-09
## 2                     age      marital_status  1.991750e-47
## 3                     age             t_stage  2.463390e-05
## 4                     age             n_stage  7.081605e-01
## 5                     age           stage_6th  1.324238e-02
## 6                     age       differentiate  4.884472e-09
## 7                     age     estrogen_status  1.477474e-04
## 8                     age               grade  4.884472e-09
## 9                     age             a_stage  1.858422e-01
## 10                    age progesterone_status  1.773628e-01
## 11                    age              status  3.866328e-04
## 12             tumor_size                race  8.960567e-01
## 13             tumor_size      marital_status  4.779921e-01
## 14             tumor_size             t_stage  0.000000e+00
## 15             tumor_size             n_stage  4.115326e-72
## 16             tumor_size           stage_6th  0.000000e+00
## 17             tumor_size       differentiate  6.452048e-13
## 18             tumor_size     estrogen_status  1.556163e-04
## 19             tumor_size               grade  6.452048e-13
## 20             tumor_size             a_stage  3.108676e-15
## 21             tumor_size progesterone_status  9.128903e-06
## 22             tumor_size              status  1.237749e-17
## 23 regional_node_examined                race  7.153662e-01
## 24 regional_node_examined      marital_status  7.987692e-01
## 25 regional_node_examined             t_stage  2.088462e-12
## 26 regional_node_examined             n_stage 1.413748e-102
## 27 regional_node_examined           stage_6th  1.164122e-99
## 28 regional_node_examined       differentiate  2.937577e-07
## 29 regional_node_examined     estrogen_status  4.445279e-03
## 30 regional_node_examined               grade  2.937577e-07
```

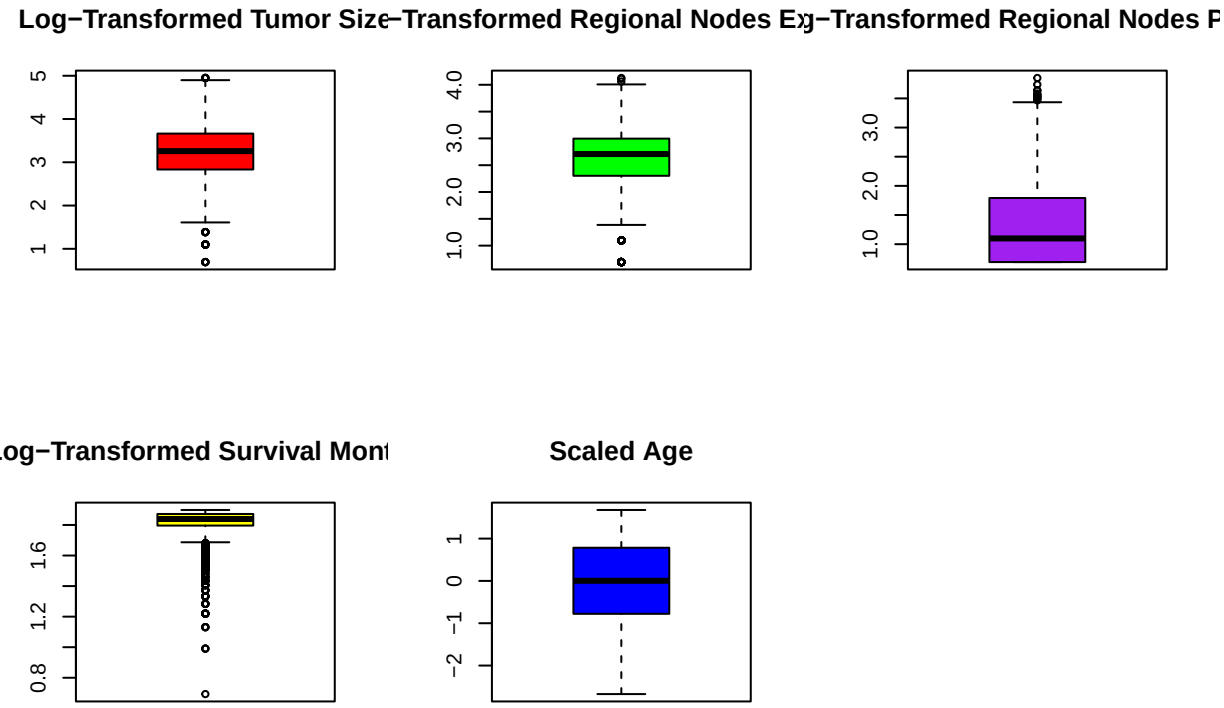
```
## 31 regional_node_examined             a_stage  1.178478e-05
## 32 regional_node_examined progesterone_status  2.522988e-01
## 33 regional_node_examined              status  2.740195e-02
## 34 regional_node_positive                race  6.361521e-01
## 35 regional_node_positive      marital_status  8.435976e-04
## 36 regional_node_positive             t_stage  7.161163e-53
## 37 regional_node_positive             n_stage  0.000000e+00
## 38 regional_node_positive           stage_6th  0.000000e+00
## 39 regional_node_positive       differentiate  2.754633e-16
## 40 regional_node_positive     estrogen_status  4.684220e-08
## 41 regional_node_positive               grade  2.754633e-16
## 42 regional_node_positive             a_stage  1.111708e-50
## 43 regional_node_positive progesterone_status  7.111004e-07
## 44 regional_node_positive              status  1.529031e-61
## 45        survival_months                race  1.124161e-03
## 46        survival_months      marital_status  2.928647e-03
## 47        survival_months             t_stage  1.689399e-06
## 48        survival_months             n_stage  9.241414e-21
## 49        survival_months           stage_6th  1.523507e-19
## 50        survival_months       differentiate  3.286748e-06
## 51        survival_months     estrogen_status  4.227475e-22
## 52        survival_months               grade  3.286748e-06
## 53        survival_months             a_stage  2.352757e-06
## 54        survival_months progesterone_status  6.896787e-14
## 55        survival_months              status 3.779502e-241
```

3. Numerical × Categorical Interactions (ANOVA Results): - Significant interactions observed between numerical and categorical variables: - Examples include age with race and t_stage, tumor_size with t_stage and status, and survival_months with status. - Decision: Include key interaction terms like age:t_stage, tumor_size:status, and survival_months:progesterone_status to improve model fit.

Step_7: Variable Transformation Section For all numerical variables:

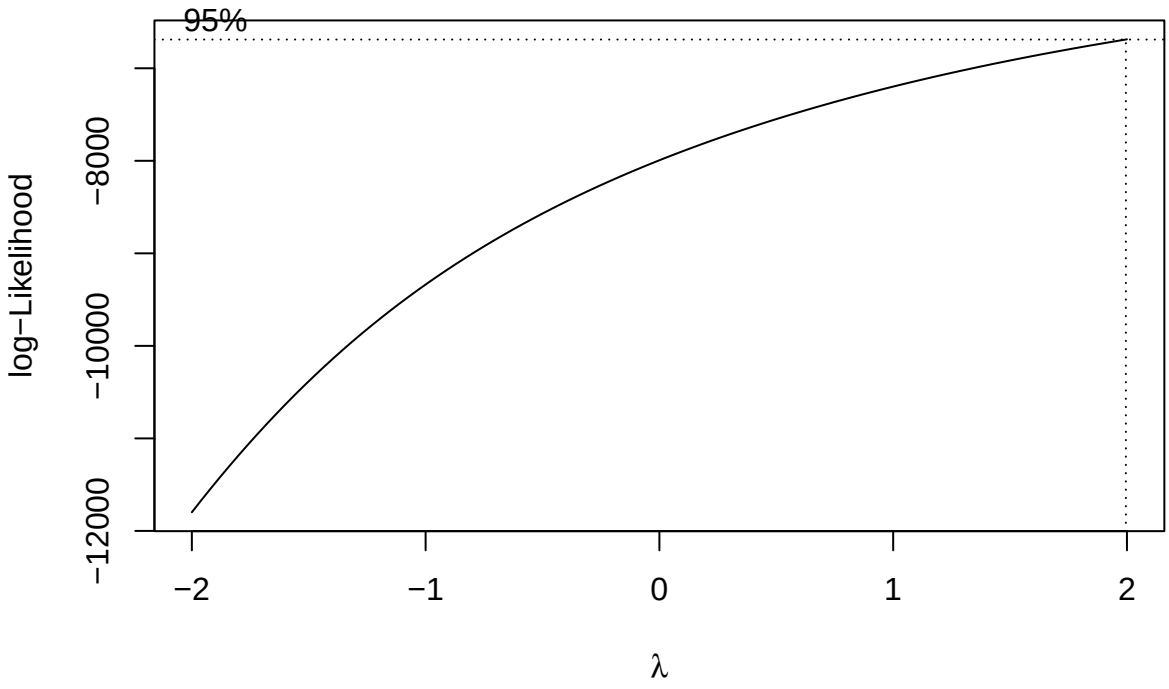
```r
# Visualize current distributions
numerical_vars <- c("age", "tumor_size", "regional_node_examined", "regional_node_positive", "survival_

# Apply transformations for skewed variables
data_clean <- data_clean %>%
  mutate(
    # Log transformation for variables with right skewness
    tumor_size_log = log(tumor_size + 1),
    regional_node_examined_log = log(regional_node_examined + 1),
    regional_node_positive_log = log(regional_node_positive + 1),
    survival_months_log = log(survival_months + 1),

    # Scale age (if necessary, based on its distribution)
    age_scaled = scale(age)
  )

# Create box plots for transformed variables
par(mfrow = c(2, 3)) # Set plotting layout
boxplot(data_clean$tumor_size_log, main = "Log-Transformed Tumor Size", col = "red")
boxplot(data_clean$regional_node_examined_log, main = "Log-Transformed Regional Nodes Examined", col =
```

```r
boxplot(data_clean$regional_node_positive_log, main = "Log-Transformed Regional Nodes Positive", col =
boxplot(data_clean$survival_months_log, main = "Log-Transformed Survival Months", col = "yellow")
boxplot(data_clean$age_scaled, main = "Scaled Age", col = "blue")
```



Further improve for survival_months (Box-Cox):

```r
# Box-Cox transformation
boxcox_result <- boxcox(lm(survival_months ~ 1, data = data_clean), lambda = seq(-2, 2, 0.1))
```




```r
# Find the best lambda
best_lambda <- boxcox_result$x[which.max(boxcox_result$y)]
cat("Best lambda:", best_lambda, "\n")
```

```
## Best lambda: 2
```

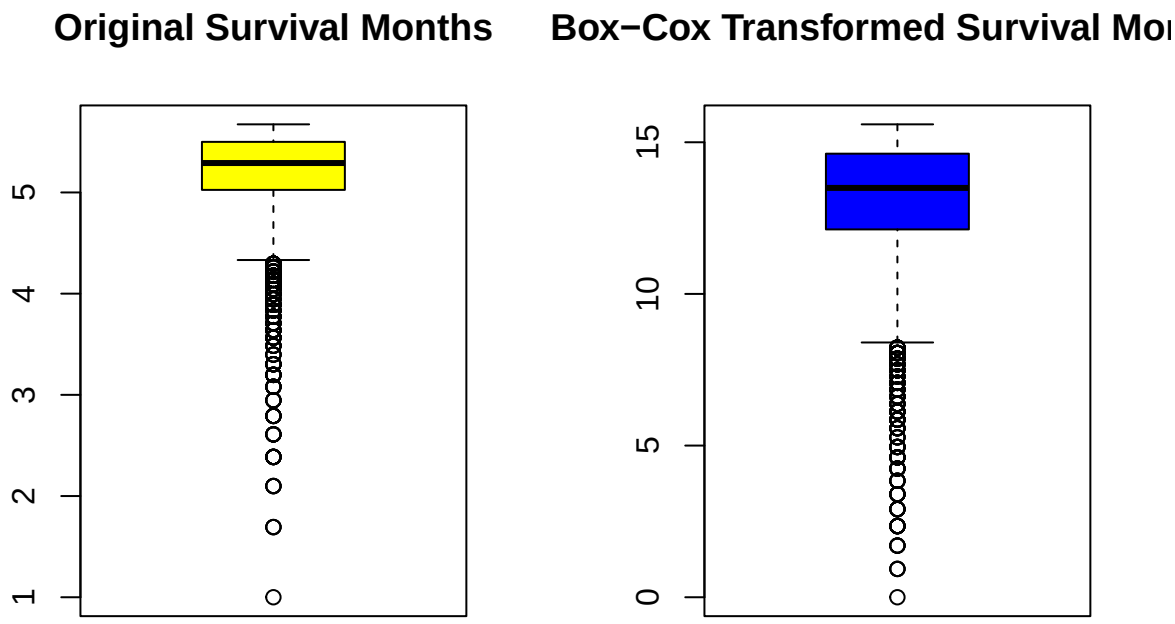
```r
# Apply the Box-Cox transformation
data_clean$survival_months_boxcox <- (data_clean$survival_months^best_lambda - 1) / best_lambda

# Check skewness before and after transformation
skewness_before <- skewness(data_clean$survival_months, na.rm = TRUE)
skewness_after <- skewness(data_clean$survival_months_boxcox, na.rm = TRUE)

cat("Skewness Before Transformation:", skewness_before, "\n")
```

```
## Skewness Before Transformation: -2.801215
```

```r
cat("Skewness After Transformation:", skewness_after, "\n")
```

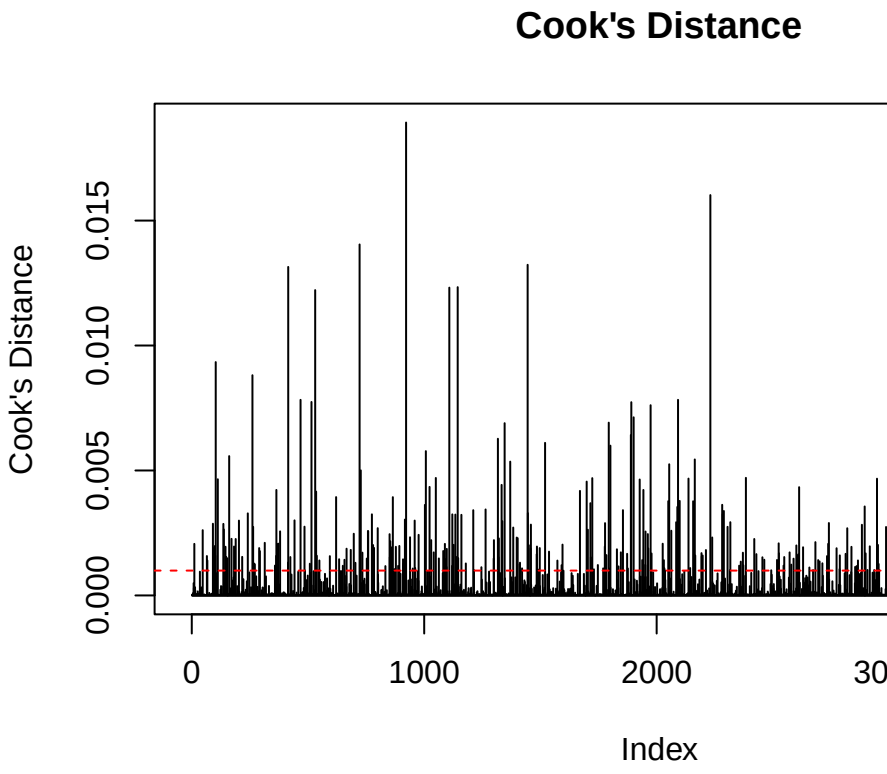
```
## Skewness After Transformation: -1.983729
```

```r
# Boxplot comparison
par(mfrow = c(1, 2))
boxplot(data_clean$survival_months, main = "Original Survival Months", col = "yellow")
boxplot(data_clean$survival_months_boxcox, main = "Box-Cox Transformed Survival Months", col = "blue")
```



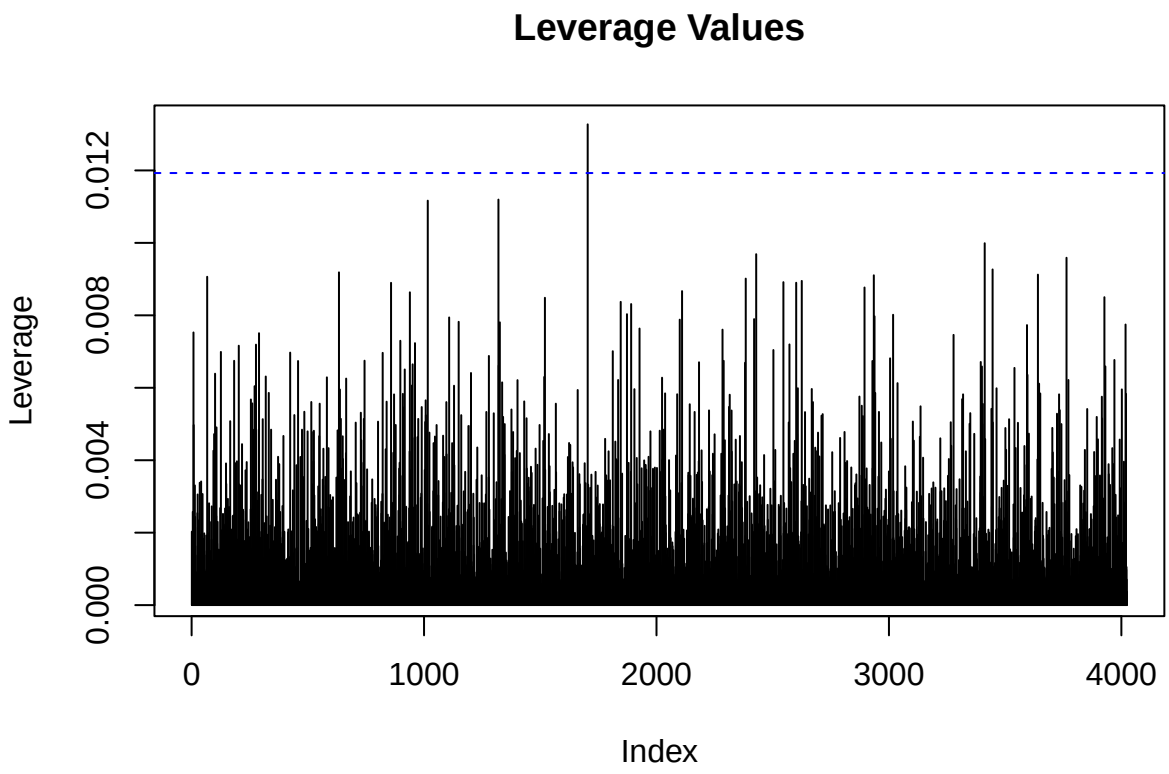

```
# Ensure your data is clean and transformed (based on previous steps)
data_clean <- data_clean %>%
  mutate(
    survival_months_boxcox = ifelse(is.na(survival_months_boxcox), survival_months, survival_months_boxc
  )

# Step 1: Fit a preliminary GLM for death risk prediction
# Assuming "status" is your binary outcome variable (1 = death, 0 = survival)
# Use a logit link function as we are predicting probabilities
glm_model <- glm(status ~ tumor_size_log + regional_node_examined_log + regional_node_positive_log +
                   age_scaled + survival_months_boxcox,
                 data = data_clean,
                 family = binomial(link = "logit"))

# Step 2: Diagnostics for Influential Points
# Leverage values (hat values)
leverage <- hatvalues(glm_model)
# Standardized residuals
std_residuals <- rstandard(glm_model)
# Cook's Distance
cooks_distance <- cooks.distance(glm_model)

# Add diagnostics back to the dataset
```

```r
data_clean <- data_clean %>%
  mutate(
    leverage = leverage,
    std_residuals = std_residuals,
    cooks_distance = cooks_distance
  )

# Step 3: Visualize Diagnostics
# Set thresholds for diagnostics
leverage_threshold <- 2 * (ncol(data_clean) - 1) / nrow(data_clean) # Rule of thumb
cooks_threshold <- 4 / nrow(data_clean) # Rule of thumb

# Plot Cook's Distance
plot(cooks_distance, type = "h", main = "Cook's Distance", ylab = "Cook's Distance", xlab = "Index")
abline(h = cooks_threshold, col = "red", lty = 2)
```

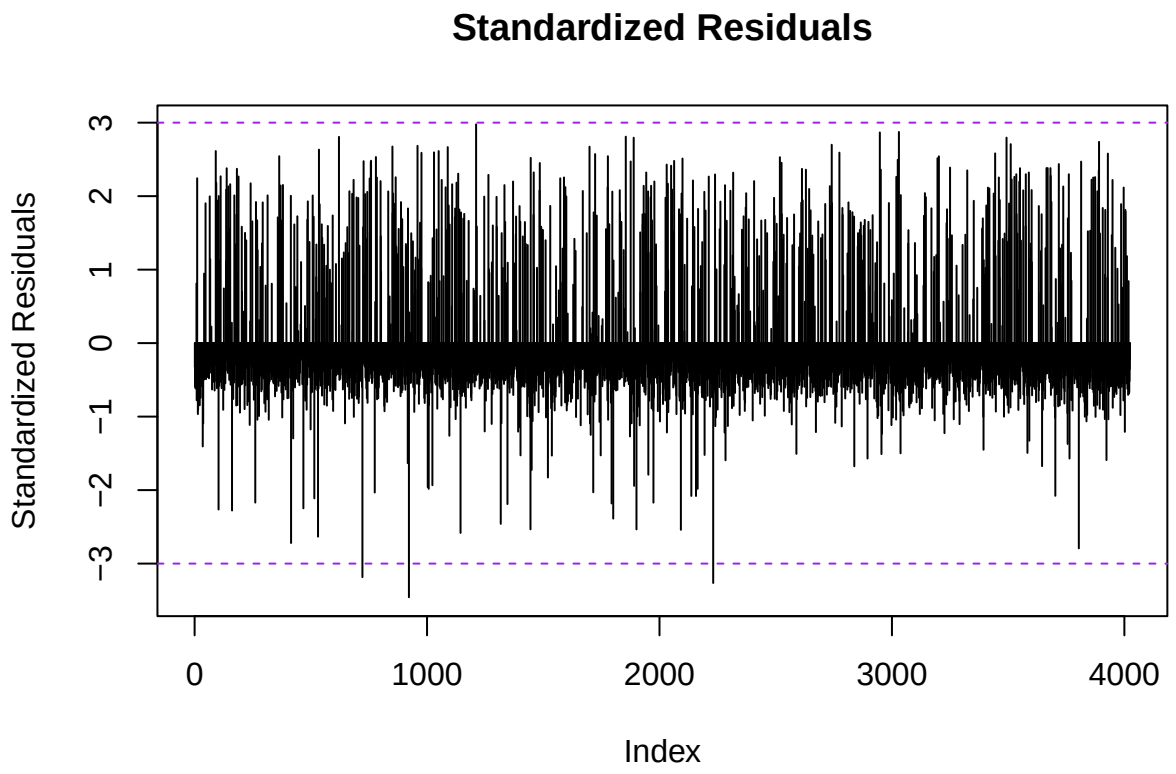


Step_8: Identifying Influencial Outliers

```r
# Plot Leverage
plot(leverage, type = "h", main = "Leverage Values", ylab = "Leverage", xlab = "Index")
abline(h = leverage_threshold, col = "blue", lty = 2)
```




```r
# Plot Standardized Residuals
plot(std_residuals, type = "h", main = "Standardized Residuals", ylab = "Standardized Residuals", xlab =
abline(h = c(-3, 3), col = "purple", lty = 2)
```




```
# Step 4: Identify Potential Influential Points
influential_points <- data_clean %>%
  filter(leverage > leverage_threshold | cooks_distance > cooks_threshold | abs(std_residuals) > 3)

# Step 5: Output Influential Points
print("Potential Influential Points:")
```

```
## [1] "Potential Influential Points:"
```

```
print(influential_points)
```

```
## # A tibble: 331 x 25
##      age race  marital_status t_stage n_stage stage_6th differentiate      grade
##    <dbl> <fct> <fct>          <fct>   <fct>   <fct>     <fct>              <fct>
##  1    68 White Widowed        T1      N1      IIA       Moderately differ~ 2
##  2    42 White Married        T1      N3      IIIC      Moderately differ~ 2
##  3    67 White Divorced       T1      N1      IIA       Moderately differ~ 2
##  4    31 White Married        T3      N3      IIIC      Poorly differenti~ 3
##  5    53 White Married        T3      N1      IIIA      Poorly differenti~ 3
##  6    63 Other Married        T2      N2      IIIA      Well differentiat~ 1
##  7    50 White Married        T4      N1      IIIB      Poorly differenti~ 3
##  8    38 White Single         T3      N3      IIIC      Moderately differ~ 2
##  9    39 White Married        T2      N1      IIB       Moderately differ~ 2
## 10    37 White Married        T3      N1      IIIA      Well differentiat~ 1
## # i 321 more rows
```

```
## # i 17 more variables: a_stage <fct>, tumor_size <dbl>, estrogen_status <fct>,
## #   progesterone_status <fct>, regional_node_examined <dbl>,
## #   regional_node_positive <dbl>, survival_months <dbl>, status <fct>,
## #   tumor_size_log <dbl>, regional_node_examined_log <dbl>,
## #   regional_node_positive_log <dbl>, survival_months_log <dbl>,
## #   age_scaled <dbl[,1]>, survival_months_boxcox <dbl>, leverage <dbl>, ...
```

```
# Summary
cat("Total influential points detected:", nrow(influential_points), "\n")
```

```
## Total influential points detected: 331
```

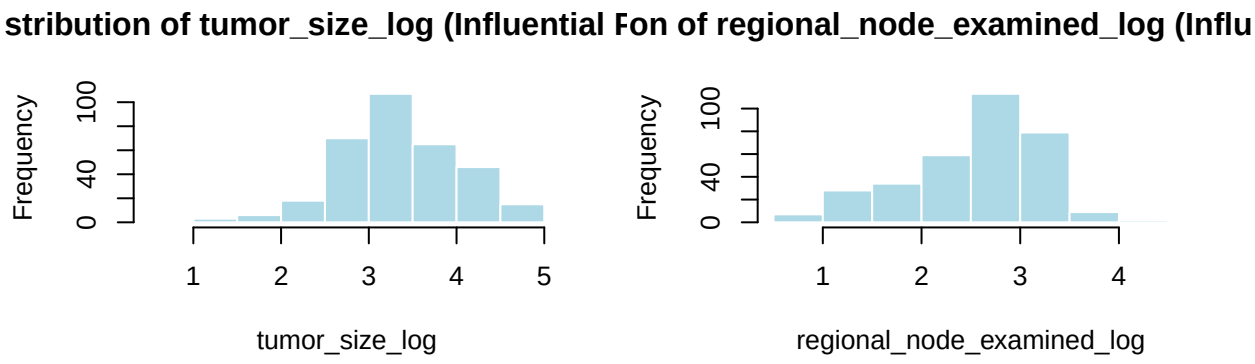
The results indicate most data points have minimal impact, but 331 potential influential points warrant further attention. Cook's Distance, leverage values, and standardized residuals highlight some points that might significantly influence the model. The next steps involve extracting these points, analyzing their distribution and origin, and deciding whether to remove errors or adjust the model to mitigate their impact. Finally, refit the model to ensure robustness.

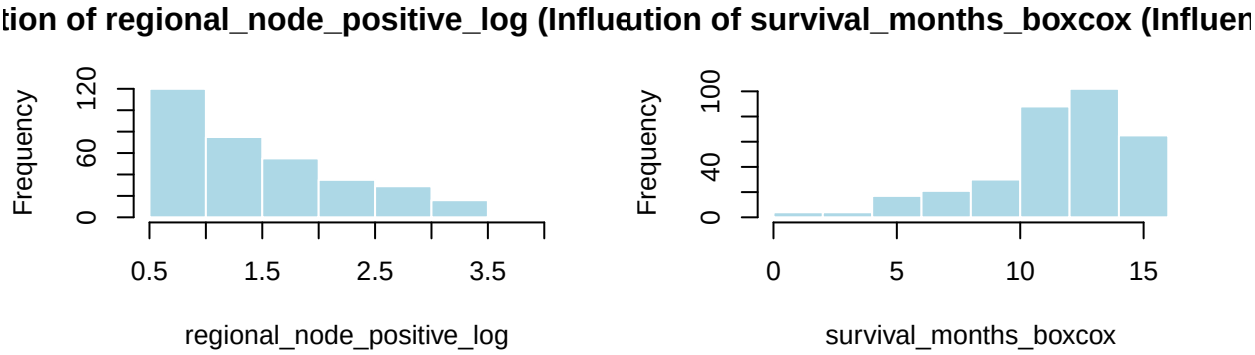
```
# Step 6: Investigate Influential Points
# Extract and analyze influential points
influential_points_data <- data_clean %>%
  filter(leverage > leverage_threshold |
         cooks_distance > cooks_threshold |
         abs(std_residuals) > 3)

# View summary of influential points
summary(influential_points_data)
```

```
##       age           race       marital_status t_stage  n_stage  stage_6th
##  Min.   :30.00   Black: 30   Divorced : 43    T1:110   N1:200   IIA :77
##  1st Qu.:45.00   Other: 27   Married  :206    T2:144   N2: 71   IIB :95
##  Median :52.00   White:274   Separated:  6    T3: 58   N3: 60   IIIA:88
##  Mean   :52.91               Single   : 55    T4: 19            IIIB:11
##  3rd Qu.:63.00               Widowed  : 21                      IIIC:60
##  Max.   :69.00
##                    differentiate                   grade         a_stage
##  Moderately differentiated:181   1                   : 30   Distant : 11
##  Poorly differentiated    :118   2                   :181   Regional:320
##  Undifferentiated         :  2   3                   :118
##  Well differentiated      : 30   anaplastic; Grade IV:  2
##
##
##    tumor_size     estrogen_status progesterone_status regional_node_examined
##  Min.   :  1.00   Negative: 32    Negative: 74        Min.   : 1.00
##  1st Qu.: 18.00   Positive:299    Positive:257        1st Qu.: 8.00
##  Median : 25.00                                       Median :14.00
##  Mean   : 34.34                                       Mean   :14.79
##  3rd Qu.: 45.00                                       3rd Qu.:20.50
##  Max.   :140.00                                       Max.   :57.00
##  regional_node_positive survival_months status  tumor_size_log
##  Min.   : 1.000         Min.   :1.000   0: 56   Min.   :0.6931
##  1st Qu.: 1.000         1st Qu.:4.638   1:275   1st Qu.:2.9444
##  Median : 2.000         Median :5.007           Median :3.2581
```

```
##  Mean   : 5.341         Mean   :4.822           Mean   :3.3462
##  3rd Qu.: 6.000         3rd Qu.:5.331           3rd Qu.:3.8286
##  Max.   :46.000         Max.   :5.625           Max.   :4.9488
##  regional_node_examined_log regional_node_positive_log survival_months_log
##  Min.   :0.6931             Min.   :0.6931             Min.   :0.6931
##  1st Qu.:2.1972             1st Qu.:0.6931             1st Qu.:1.7295
##  Median :2.7081             Median :1.0986             Median :1.7930
##  Mean   :2.5574             Mean   :1.4735             Mean   :1.7521
##  3rd Qu.:3.0678             3rd Qu.:1.9459             3rd Qu.:1.8454
##  Max.   :4.0604             Max.   :3.8501             Max.   :1.8908
##     age_scaled.V1     survival_months_boxcox    leverage
##  Min.   :-2.6745295   Min.   : 0.00          Min.   :0.0002467
##  1st Qu.:-1.0010077   1st Qu.:10.25          1st Qu.:0.0011756
##  Median :-0.2200310   Median :12.04          Median :0.0023401
##  Mean   :-0.1189118   Mean   :11.39          Mean   :0.0030386
##  3rd Qu.: 1.0072183   3rd Qu.:13.71          3rd Qu.:0.0042786
##  Max.   : 1.6766270   Max.   :15.32          Max.   :0.0132721
##  std_residuals    cooks_distance
##  Min.   :-3.458   Min.   :0.001001
##  1st Qu.: 1.391   1st Qu.:0.001413
##  Median : 1.813   Median :0.001948
##  Mean   : 1.295   Mean   :0.002725
##  3rd Qu.: 2.217   3rd Qu.:0.003001
##  Max.   : 2.977   Max.   :0.018923
```

```r
# Visualize distributions for key numeric variables
# Specify the variables of interest
key_numeric_vars <- c("tumor_size_log", "regional_node_examined_log",
                      "regional_node_positive_log", "survival_months_boxcox")

# Set up the plotting layout
par(mfrow = c(2, 2))  # Arrange plots in a 2x2 grid
for (var in key_numeric_vars) {
  if (var %in% colnames(influential_points_data)) {
    hist(influential_points_data[[var]],
         main = paste("Distribution of", var, "(Influential Points)"),
         xlab = var, col = "lightblue", border = "white")
  }
}
```






```r
# Reset plotting layout to default
par(mfrow = c(1, 1))

# Step 7: Handle Influential Points
# Exclude influential points (basic cleaning for modeling preparation)
data_clean_no_outliers <- data_clean %>%
  filter(!(leverage > leverage_threshold |
           cooks_distance > cooks_threshold |
           abs(std_residuals) > 3))

# Refit the GLM model to ensure data is ready for modeling
glm_model_updated <- glm(status ~ tumor_size_log + regional_node_examined_log +
                         regional_node_positive_log + age_scaled +
                         survival_months_boxcox,
                         data = data_clean_no_outliers,
                         family = binomial(link = "logit"))

# Evaluate the updated model
cat("Summary of the updated model:\n")
```

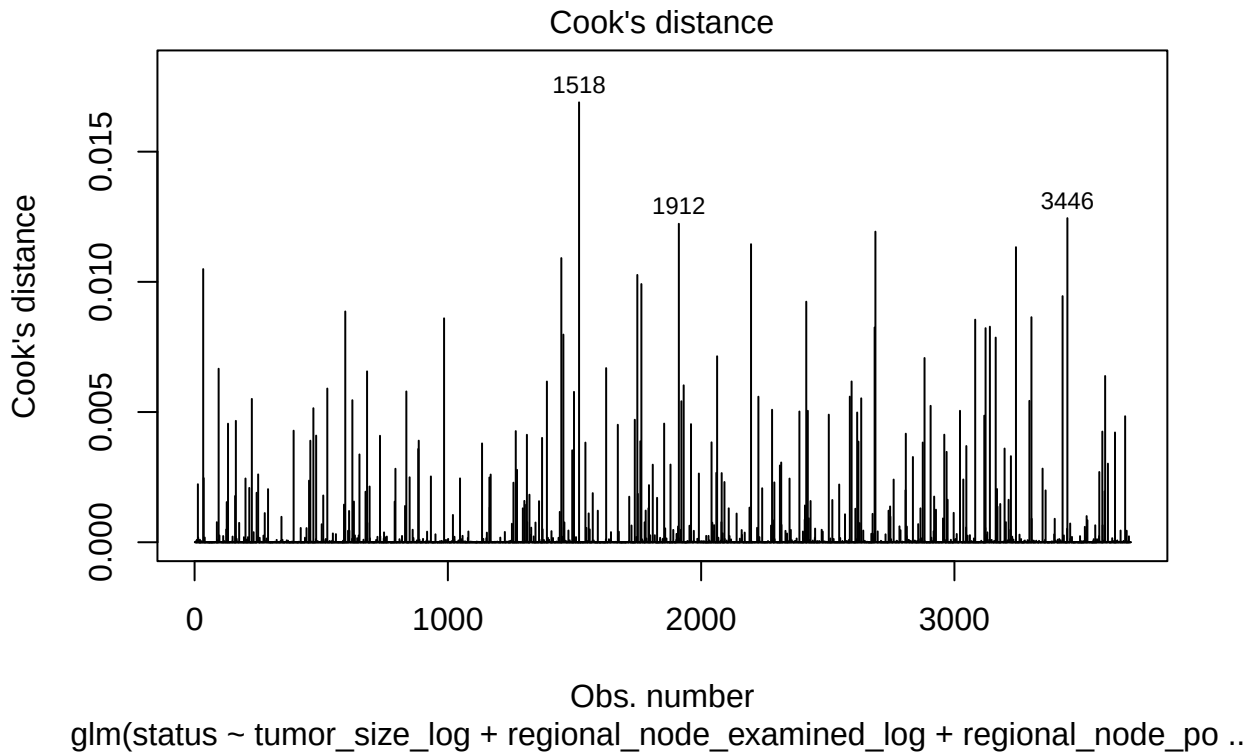
```
## Summary of the updated model:
```

```r
summary(glm_model_updated)
```

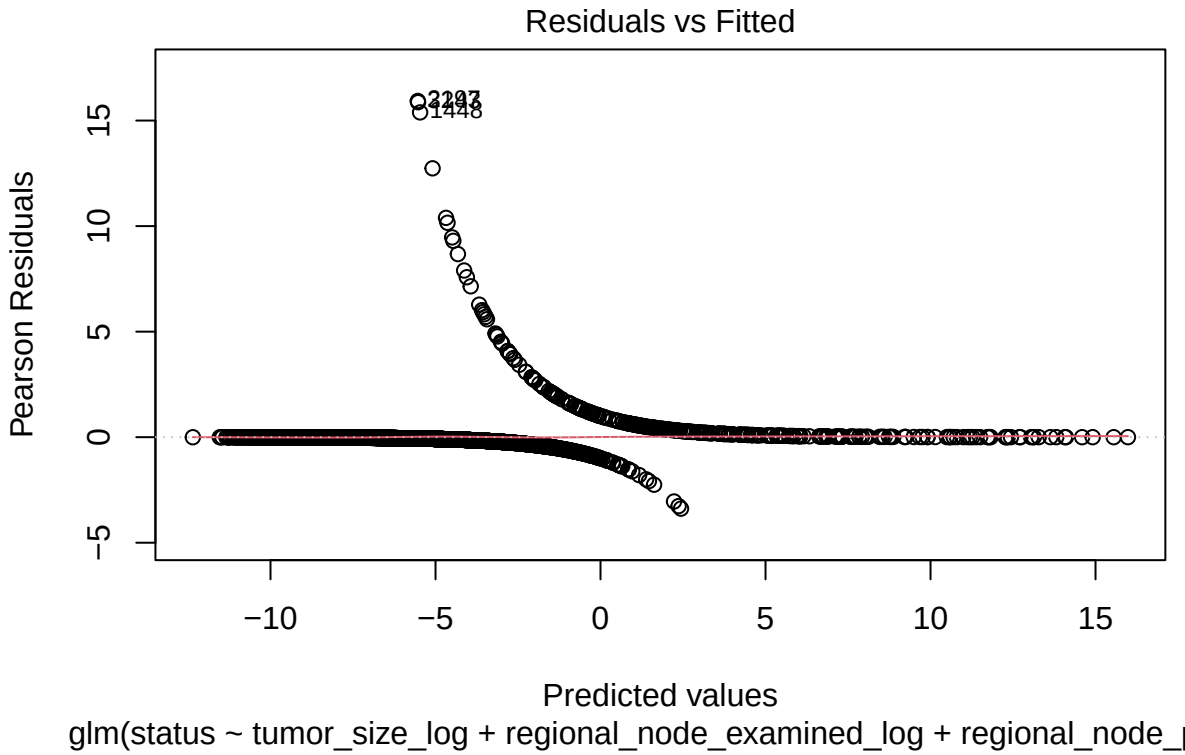
```
##
```

```
## Call:
## glm(formula = status ~ tumor_size_log + regional_node_examined_log +
##     regional_node_positive_log + age_scaled + survival_months_boxcox,
##     family = binomial(link = "logit"), data = data_clean_no_outliers)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  13.50503    1.21919  11.077  < 2e-16 ***
## tumor_size_log                0.62791    0.18550   3.385 0.000712 ***
## regional_node_examined_log -0.71026    0.22420  -3.168 0.001535 **
## regional_node_positive_log  1.82088    0.18275   9.964  < 2e-16 ***
## age_scaled                    0.51745    0.11806   4.383 1.17e-05 ***
## survival_months_boxcox     -1.61551    0.09271 -17.425  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2274.2  on 3692  degrees of freedom
## Residual deviance:  663.3  on 3687  degrees of freedom
## AIC: 675.3
##
## Number of Fisher Scoring iterations: 8
```

```r
# Diagnostic plots (optional)
plot(glm_model_updated, which = 4)  # Cook's Distance
```



```r
plot(glm_model_updated, which = 1)  # Residuals vs Fitted
```




```r
# Final Message
cat("Process completed. Key numeric variables visualized, updated model fitted, and cleaned dataset sav
```

```
## Process completed. Key numeric variables visualized, updated model fitted, and cleaned dataset saved
```

Comments on Outputs - The cleaned model performs well, with all variables being statistically significant. The residual deviance has significantly decreased (663.3), and the AIC value (675.3) indicates a good balance between model fit and complexity. Diagnostic plots confirm that there are no significant high-influence points remaining, and the distributions of key variables are reasonable without severe outliers or biases, demonstrating the effectiveness of the data cleaning process.

Final Decision - We have decided to remove the 300+ influential outliers identified earlier and proceed with the cleaned dataset data_clean_no_outliers as the foundation for modeling. The data quality is sufficient, marking the conclusion of the data exploration phase and the transition to the modeling stage.

II. Modeling

```r
# Ensure data balance
# Split data based on status
data_majority <- data_clean_no_outliers %>% filter(status == 0)
```

```r
data_minority <- data_clean_no_outliers %>% filter(status == 1)
set.seed(123)
data_majority_sample <- data_majority %>% sample_n(min(nrow(data_minority) * 2, nrow(data_majority)))
data_balanced <- bind_rows(data_majority_sample, data_minority)

# Standardize numerical variables
data_balanced$tumor_size_log <- scale(data_balanced$tumor_size_log)
data_balanced$regional_node_examined_log <- scale(data_balanced$regional_node_examined_log)
data_balanced$regional_node_positive_log <- scale(data_balanced$regional_node_positive_log)
data_balanced$survival_months_boxcox <- scale(data_balanced$survival_months_boxcox)
data_balanced$age_scaled <- scale(data_balanced$age_scaled)

# Split data into training and testing sets
set.seed(123)
trainIndex <- createDataPartition(data_balanced$status, p = 0.8, list = FALSE)
train_data <- data_balanced[trainIndex, ]
test_data <- data_balanced[-trainIndex, ]

# Prepare training data for Lasso regression
x_train <- model.matrix(status ~ tumor_size_log + regional_node_examined_log +
                        regional_node_positive_log + survival_months_boxcox +
                        age_scaled + race + marital_status + t_stage + n_stage +
                        stage_6th + differentiate + estrogen_status + grade +
                        a_stage + progesterone_status + age_scaled:t_stage +
                        tumor_size_log:status + survival_months_boxcox:progesterone_status,
                        data = train_data)[, -1]
y_train <- train_data$status

# Perform Lasso regression
lasso_model <- cv.glmnet(x_train, y_train, alpha = 1, family = "binomial")
best_lambda <- lasso_model$lambda.min
model_lasso <- glmnet(x_train, y_train, alpha = 1, lambda = best_lambda, family = "binomial")

# Extract coefficients from the Lasso model
coef_lasso <- coef(model_lasso, s = best_lambda)
selected_variables_df <- as.data.frame(as.matrix(coef_lasso))
colnames(selected_variables_df) <- "Coefficient"
selected_variables_df <- selected_variables_df[selected_variables_df$Coefficient != 0, , drop = FALSE]
selected_variables_df <- cbind(Variable = rownames(selected_variables_df), selected_variables_df)
rownames(selected_variables_df) <- NULL
print("Selected variables by Lasso model:")
```

Step_1: Logistic Regression Model with Interaction Terms and Optimization

```
## [1] "Selected variables by Lasso model:"
```

```r
# Present coefficients in a table format
knitr::kable(selected_variables_df, caption = "Coefficients of the Optimal Lasso Model")
```

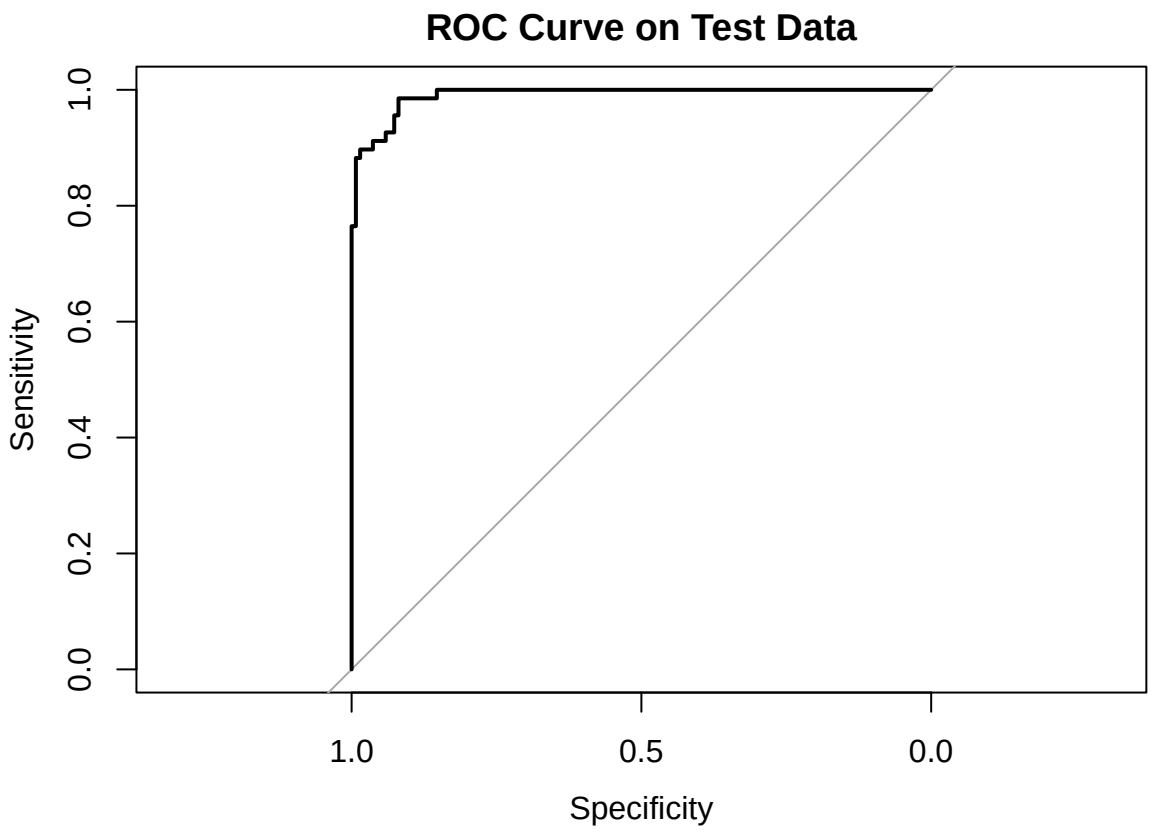
Table 3: Coefficients of the Optimal Lasso Model

Variable	Coefficient
(Intercept)	-1.1126229
regional_node_examined_log	-0.2127777
regional_node_positive_log	1.2289846
survival_months_boxcox	-4.2650954
age_scaled	0.1765491
raceOther	-1.7033045
raceWhite	-0.5564617
marital_statusSingle	-0.0955304
marital_statusWidowed	-0.0098700
t_stageT2	0.9403077
n_stageN3	0.0913812
stage_6thIIIC	0.0514802
differentiatePoorly differentiated	0.6297549
differentiateUndifferentiated	3.3406388
differentiateWell differentiated	-0.1250715
estrogen_statusPositive	0.1671107
grade3	0.0239706
gradeanaplastic; Grade IV	0.0118946
progesterone_statusPositive	-0.6840958
age_scaled:t_stageT4	0.1264922
status1:tumor_size_log	0.5409229

```
# Validate the model on the testing set
x_test <- model.matrix(status ~ tumor_size_log + regional_node_examined_log +
                       regional_node_positive_log + survival_months_boxcox +
                       age_scaled + race + marital_status + t_stage + n_stage +
                       stage_6th + differentiate + estrogen_status + grade +
                       a_stage + progesterone_status + age_scaled:t_stage +
                       tumor_size_log:status + survival_months_boxcox:progesterone_status,
                       data = test_data)[, -1]
y_test <- test_data$status
test_predictions <- as.numeric(predict(model_lasso, newx = x_test, type = "response"))

# Calculate AUC on the testing set
roc_curve_test <- roc(y_test, test_predictions)
auc_test <- auc(roc_curve_test)
print(paste("Test AUC:", auc_test))
```

```
## [1] "Test AUC: 0.990808823529412"
```

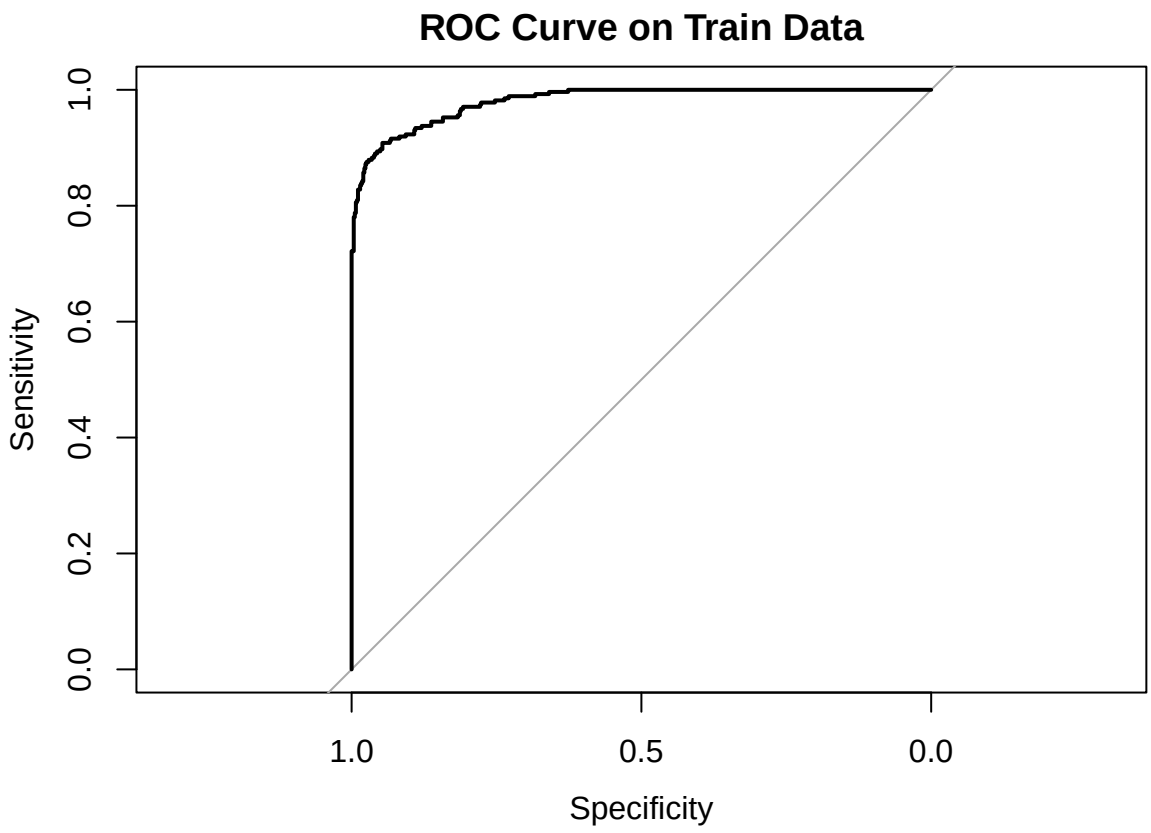
```
plot(roc_curve_test, main = "ROC Curve on Test Data")
```




```
# Calculate AUC on the training set
train_predictions <- as.numeric(predict(model_lasso, newx = x_train, type = "response"))
roc_curve_train <- roc(y_train, train_predictions)
auc_train <- auc(roc_curve_train)
print(paste("Train AUC:", auc_train))
```

```
## [1] "Train AUC: 0.980329804505626"
```

```
plot(roc_curve_train, main = "ROC Curve on Train Data")
```




```
# Data preparation
cox_data <- data_clean_no_outliers
cox_data$status <- as.numeric(cox_data$status)  # Cox model requires status to be numeric (0: alive, 1:

# Define the Cox model formula, including main effects and interaction terms
cox_formula <- as.formula("Surv(survival_months_boxcox, status) ~
                          tumor_size_log + regional_node_examined_log +
                          regional_node_positive_log + age_scaled + race +
                          marital_status + t_stage + n_stage + stage_6th +
                          differentiate + estrogen_status + grade + a_stage +
                          progesterone_status + age_scaled:t_stage +
                          tumor_size_log:status + survival_months_boxcox:progesterone_status")

# Fit the Cox proportional hazards model
cox_model <- coxph(cox_formula, data = cox_data)

# Output model results
summary(cox_model)
```

Step_2: Cox Proportional Hazards Model

```
## Call:
## coxph(formula = cox_formula, data = cox_data)
```

```
##
##   n= 3693, number of events= 341
##
##                                                       coef  exp(coef)
## tumor_size_log                                   -8.986e+00  1.251e-04
## regional_node_examined_log                       -4.719e-05  1.000e+00
## regional_node_positive_log                        4.788e-04  1.000e+00
## age_scaled                                        1.748e-03  1.002e+00
## raceOther                                         3.291e-03  1.003e+00
## raceWhite                                         4.189e-03  1.004e+00
## marital_statusMarried                             8.754e-04  1.001e+00
## marital_statusSeparated                          -6.155e-05  9.999e-01
## marital_statusSingle                              1.460e-03  1.001e+00
## marital_statusWidowed                             2.276e-03  1.002e+00
## t_stageT2                                        -1.013e-02  9.899e-01
## t_stageT3                                        -1.818e-02  9.820e-01
## t_stageT4                                        -1.025e-02  9.898e-01
## n_stageN2                                         2.137e-03  1.002e+00
## n_stageN3                                         3.567e-03  1.004e+00
## stage_6thIIB                                      7.061e-03  1.007e+00
## stage_6thIIIA                                     3.266e-03  1.003e+00
## stage_6thIIIB                                    -3.941e-03  9.961e-01
## stage_6thIIIC                                     0.000e+00  1.000e+00
## differentiatePoorly differentiated                2.106e-03  1.002e+00
## differentiateUndifferentiated                    -7.354e-04  9.993e-01
## differentiateWell differentiated                  4.621e-03  1.005e+00
## estrogen_statusPositive                           3.077e-03  1.003e+00
## grade2                                            0.000e+00  1.000e+00
## grade3                                            0.000e+00  1.000e+00
## gradeanaplastic; Grade IV                         0.000e+00  1.000e+00
## a_stageRegional                                   9.840e-04  1.001e+00
## progesterone_statusPositive                      -2.068e-03  9.979e-01
## age_scaled:t_stageT2                             -2.288e-03  9.977e-01
## age_scaled:t_stageT3                             -3.805e-03  9.962e-01
## age_scaled:t_stageT4                             -4.166e-03  9.958e-01
## tumor_size_log:status                             4.500e+00  9.001e+01
## progesterone_statusNegative:survival_months_boxcox -5.528e+01  9.867e-25
## progesterone_statusPositive:survival_months_boxcox -5.528e+01  9.867e-25
##                                                    se(coef)         z
## tumor_size_log                                     1.055e-01   -85.182
## regional_node_examined_log                         1.086e-01     0.000
## regional_node_positive_log                         7.858e-02     0.006
## age_scaled                                         6.338e-02     0.028
## raceOther                                          2.964e-01     0.011
## raceWhite                                          1.608e-01     0.026
## marital_statusMarried                              1.210e-01     0.007
## marital_statusSeparated                            3.516e-01     0.000
## marital_statusSingle                               1.607e-01     0.009
## marital_statusWidowed                              2.098e-01     0.011
## t_stageT2                                          1.220e-01    -0.083
## t_stageT3                                          1.521e-01    -0.119
## t_stageT4                                          2.438e-01    -0.042
## n_stageN2                                          1.292e-01     0.017
## n_stageN3                                          1.237e-01     0.029
```

```
## stage_6thIIB                                          1.598e-01     0.044
## stage_6thIIIA                                         1.269e-01     0.026
## stage_6thIIIB                                         3.786e-01    -0.010
## stage_6thIIIC                                         1.237e-01     0.000
## differentiatePoorly differentiated                    1.244e-01     0.017
## differentiateUndifferentiated                         4.478e-01    -0.002
## differentiateWell differentiated                      2.791e-01     0.017
## estrogen_statusPositive                               1.521e-01     0.020
## grade2                                                1.259e-01     0.000
## grade3                                                1.244e-01     0.000
## gradeanaplastic; Grade IV                             4.478e-01     0.000
## a_stageRegional                                       2.310e-01     0.004
## progesterone_statusPositive                           1.268e-01    -0.016
## age_scaled:t_stageT2                                  8.594e-02    -0.027
## age_scaled:t_stageT3                                  1.438e-01    -0.026
## age_scaled:t_stageT4                                  2.130e-01    -0.020
## tumor_size_log:status                                 5.259e-02    85.566
## progesterone_statusNegative:survival_months_boxcox  1.411e-02 -3918.242
## progesterone_statusPositive:survival_months_boxcox  1.411e-02 -3918.248
##                                                      Pr(>|z|)
## tumor_size_log                                         <2e-16 ***
## regional_node_examined_log                              1.000
## regional_node_positive_log                              0.995
## age_scaled                                              0.978
## raceOther                                               0.991
## raceWhite                                               0.979
## marital_statusMarried                                   0.994
## marital_statusSeparated                                 1.000
## marital_statusSingle                                    0.993
## marital_statusWidowed                                   0.991
## t_stageT2                                               0.934
## t_stageT3                                               0.905
## t_stageT4                                               0.966
## n_stageN2                                               0.987
## n_stageN3                                               0.977
## stage_6thIIB                                            0.965
## stage_6thIIIA                                           0.979
## stage_6thIIIB                                           0.992
## stage_6thIIIC                                           1.000
## differentiatePoorly differentiated                      0.986
## differentiateUndifferentiated                           0.999
## differentiateWell differentiated                        0.987
## estrogen_statusPositive                                 0.984
## grade2                                                  1.000
## grade3                                                  1.000
## gradeanaplastic; Grade IV                               1.000
## a_stageRegional                                         0.997
## progesterone_statusPositive                             0.987
## age_scaled:t_stageT2                                    0.979
## age_scaled:t_stageT3                                    0.979
## age_scaled:t_stageT4                                    0.984
## tumor_size_log:status                                  <2e-16 ***
## progesterone_statusNegative:survival_months_boxcox     <2e-16 ***
## progesterone_statusPositive:survival_months_boxcox     <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                    exp(coef) exp(-coef)
## tumor_size_log                                     1.251e-04  7.993e+03
## regional_node_examined_log                         1.000e+00  1.000e+00
## regional_node_positive_log                         1.000e+00  9.995e-01
## age_scaled                                         1.002e+00  9.983e-01
## raceOther                                          1.003e+00  9.967e-01
## raceWhite                                          1.004e+00  9.958e-01
## marital_statusMarried                              1.001e+00  9.991e-01
## marital_statusSeparated                            9.999e-01  1.000e+00
## marital_statusSingle                               1.001e+00  9.985e-01
## marital_statusWidowed                              1.002e+00  9.977e-01
## t_stageT2                                          9.899e-01  1.010e+00
## t_stageT3                                          9.820e-01  1.018e+00
## t_stageT4                                          9.898e-01  1.010e+00
## n_stageN2                                          1.002e+00  9.979e-01
## n_stageN3                                          1.004e+00  9.964e-01
## stage_6thIIB                                       1.007e+00  9.930e-01
## stage_6thIIIA                                      1.003e+00  9.967e-01
## stage_6thIIIB                                      9.961e-01  1.004e+00
## stage_6thIIIC                                      1.000e+00  1.000e+00
## differentiatePoorly differentiated                 1.002e+00  9.979e-01
## differentiateUndifferentiated                      9.993e-01  1.001e+00
## differentiateWell differentiated                   1.005e+00  9.954e-01
## estrogen_statusPositive                            1.003e+00  9.969e-01
## grade2                                             1.000e+00  1.000e+00
## grade3                                             1.000e+00  1.000e+00
## gradeanaplastic; Grade IV                          1.000e+00  1.000e+00
## a_stageRegional                                    1.001e+00  9.990e-01
## progesterone_statusPositive                        9.979e-01  1.002e+00
## age_scaled:t_stageT2                               9.977e-01  1.002e+00
## age_scaled:t_stageT3                               9.962e-01  1.004e+00
## age_scaled:t_stageT4                               9.958e-01  1.004e+00
## tumor_size_log:status                              9.001e+01  1.111e-02
## progesterone_statusNegative:survival_months_boxcox 9.867e-25  1.014e+24
## progesterone_statusPositive:survival_months_boxcox 9.867e-25  1.014e+24
##                                                    lower .95 upper .95
## tumor_size_log                                     1.017e-04 1.539e-04
## regional_node_examined_log                         8.083e-01 1.237e+00
## regional_node_positive_log                         8.577e-01 1.167e+00
## age_scaled                                         8.847e-01 1.134e+00
## raceOther                                          5.613e-01 1.793e+00
## raceWhite                                          7.327e-01 1.376e+00
## marital_statusMarried                              7.896e-01 1.269e+00
## marital_statusSeparated                            5.019e-01 1.992e+00
## marital_statusSingle                               7.308e-01 1.372e+00
## marital_statusWidowed                              6.644e-01 1.512e+00
## t_stageT2                                          7.794e-01 1.257e+00
## t_stageT3                                          7.288e-01 1.323e+00
## t_stageT4                                          6.138e-01 1.596e+00
## n_stageN2                                          7.779e-01 1.291e+00
## n_stageN3                                          7.876e-01 1.279e+00
```

```
## stage_6thIIB                                          7.363e-01 1.378e+00
## stage_6thIIIA                                         7.823e-01 1.287e+00
## stage_6thIIIB                                         4.743e-01 2.092e+00
## stage_6thIIIC                                         7.848e-01 1.274e+00
## differentiatePoorly differentiated                    7.852e-01 1.279e+00
## differentiateUndifferentiated                         4.154e-01 2.404e+00
## differentiateWell differentiated                      5.813e-01 1.736e+00
## estrogen_statusPositive                               7.446e-01 1.351e+00
## grade2                                                7.814e-01 1.280e+00
## grade3                                                7.836e-01 1.276e+00
## gradeanaplastic; Grade IV                             4.157e-01 2.405e+00
## a_stageRegional                                       6.365e-01 1.574e+00
## progesterone_statusPositive                           7.783e-01 1.280e+00
## age_scaled:t_stageT2                                  8.430e-01 1.181e+00
## age_scaled:t_stageT3                                  7.515e-01 1.321e+00
## age_scaled:t_stageT4                                  6.560e-01 1.512e+00
## tumor_size_log:status                                 8.120e+01 9.978e+01
## progesterone_statusNegative:survival_months_boxcox 9.597e-25 1.014e-24
## progesterone_statusPositive:survival_months_boxcox 9.598e-25 1.014e-24
##
## Concordance= 1  (se = 0 )
## Likelihood ratio test= 4792  on 34 df,   p=<2e-16
## Wald test            = 30719868  on 34 df,   p=<2e-16
## Score (logrank) test = 7602  on 34 df,   p=<2e-16
```

```
# Check model performance - calculate C-index
cindex <- summary(cox_model)$concordance[1]
print(paste("C-index:", cindex))
```

```
## [1] "C-index: 1"
```

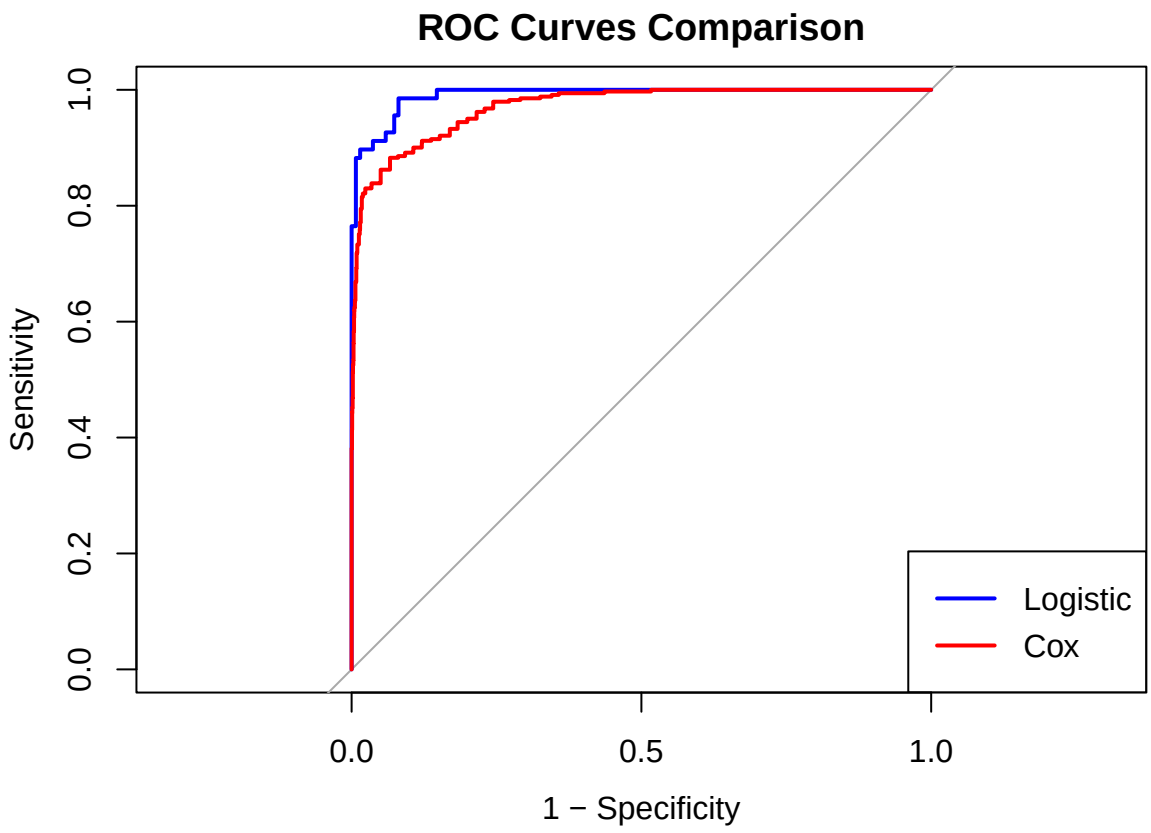
```
# Compare model prediction performance
# Predictions from the logistic model (assuming test_predictions are generated)
roc_logistic <- roc(y_test, test_predictions)
auc_logistic <- auc(roc_logistic)
print(paste("Logistic Test AUC:", auc_logistic))
```

```
## [1] "Logistic Test AUC: 0.990808823529412"
```

```
# Risk scores from the Cox model
cox_risk <- predict(cox_model, type = "risk")
roc_cox <- roc(cox_data$status, cox_risk)
auc_cox <- auc(roc_cox)
print(paste("Cox Model AUC:", auc_cox))
```

```
## [1] "Cox Model AUC: 0.970237053730773"
```

```
# Output ROC curve comparison
plot(roc_logistic, main = "ROC Curves Comparison", col = "blue", legacy.axes = TRUE)
lines(roc_cox, col = "red")
legend("bottomright", legend = c("Logistic", "Cox"), col = c("blue", "red"), lwd = 2)
```




```
# Extract coefficients from the Cox model
cox_summary <- summary(cox_model)
coefficients <- as.data.frame(cox_summary$coefficients)
colnames(coefficients) <- c("Coefficient", "Exp(Coefficient)", "Standard Error", "z-value", "p-value")

# Create and visualize the table for all coefficients
coefficients_table <- coefficients %>%
  rownames_to_column(var = "Variable") %>%
  kable("html", caption = "Cox Model Coefficients") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"), full_width = FALSE

# Show the table for all coefficients
coefficients_table
```

Cox Model Coefficients

Variable

Coefficient

Exp(Coefficient)

Standard Error

z-value

p-value

tumor_size_log

-8.9862884

0.0001251

0.1054954

-85.1817826

0.0000000

regional_node_examined_log

-0.0000472

0.9999528

0.1085778

-0.0004346

0.9996532

regional_node_positive_log

0.0004788

1.0004789

0.0785832

0.0060927

0.9951387

age_scaled

0.0017484

1.0017499

0.0633788

0.0275863

0.9779921

raceOther

0.0032910

1.0032964

0.2963668

0.0111045

0.9911401

raceWhite

0.0041888

1.0041975

0.1608252

0.0260455

0.9792211

marital_statusMarried

0.0008754

1.0008758

0.1209759

0.0072363

0.9942263

marital_statusSeparated

-0.0000615

0.9999385

0.3516405

-0.0001750

0.9998603

marital_statusSingle

0.0014597

1.0014607

0.1607326

0.0090813

0.9927542

marital_statusWidowed

0.0022764

1.0022790

0.2098098

0.0108499

0.9913432

t_stageT2

-0.0101315

0.9899197

0.1220186

-0.0830320

0.9338261

t_stageT3

-0.0181759

0.9819883

0.1521144

-0.1194885

0.9048884

t_stageT4

-0.0102453

0.9898070

0.2438168

-0.0420204

0.9664824

n_stageN2

0.0021370

1.0021393

0.1292166

0.0165380

0.9868052

n_stageN3

0.0035670

1.0035733

0.1236519

0.0288468

0.9769868

stage_6thIIB

0.0070613

1.0070863

0.1598149

0.0441841

0.9647577

stage_6thIIIA

0.0032658

1.0032712

0.1269499

0.0257255

0.9794763

stage_6thIIIB

-0.0039414

0.9960664

0.3785811

-0.0104109

0.9916934

stage_6thIIIC

0.0000000

1.0000000

0.1236519

0.0000000

1.0000000

differentiatePoorly differentiated

0.0021058

1.0021080

0.1244206

0.0169247

0.9864967

differentiateUndifferentiated

-0.0007354

0.9992649

0.4477999

-0.0016422

0.9986897

differentiateWell differentiated

0.0046211

1.0046318

0.2791066

0.0165569

0.9867901

estrogen_statusPositive

0.0030773

1.0030820

0.1520626

0.0202371

0.9838542

grade2

0.0000000

1.0000000

0.1258793

0.0000000

1.0000000

grade3

0.0000000

1.0000000

0.1244206

0.0000000

1.0000000

gradeanaplastic; Grade IV

0.0000000

1.0000000

0.4477999

0.0000000

1.0000000

a_stageRegional

0.0009840

1.0009845

0.2310265

0.0042594

0.9966015

progesterone_statusPositive

-0.0020682

0.9979340

0.1268311

-0.0163065

0.9869898

age_scaled:t_stageT2

-0.0022880

0.9977146

0.0859432

-0.0266223

0.9787610

age_scaled:t_stageT3

-0.0038048

0.9962025

0.1438325

-0.0264527

0.9788963

age_scaled:t_stageT4

-0.0041662

0.9958425

0.2129544

-0.0195637

0.9843914

tumor_size_log:status

4.4999286

90.0107010

0.0525902

85.5659521

0.0000000

progesterone_statusNegative:survival_months_boxcox

-55.2754780

0.0000000

0.0141072

-3918.2418706

0.0000000

progesterone_statusPositive:survival_months_boxcox

-55.2754619

0.0000000

0.0141072

-3918.2484653

0.0000000

```
# Create and visualize the table for significant coefficients (p-value < 0.05)
significant_coefficients <- coefficients %>%
  filter(`p-value` < 0.05) %>%
  rownames_to_column(var = "Variable") %>%
  kable("html", caption = "Significant Cox Model Coefficients") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"), full_width = FALSE

# Show the table for significant coefficients
significant_coefficients
```

Significant Cox Model Coefficients

Variable

Coefficient

Exp(Coefficient)

Standard Error

z-value

p-value

tumor_size_log

-8.986288

0.0001251

0.1054954

-85.18178

0

tumor_size_log:status

4.499929

90.0107010

0.0525902

85.56595

0

progesterone_statusNegative:survival_months_boxcox

-55.275478

0.0000000

0.0141072

-3918.24187

0

progesterone_statusPositive:survival_months_boxcox

-55.275462

0.0000000

0.0141072

-3918.24847

0

```
# Calculate model performance for different racial groups
# Logistic model - Compute AUC by race
roc_logistic_white <- roc(
  test_data %>% filter(race == "White") %>% pull(status),
  test_predictions[test_data$race == "White"]
)

roc_logistic_black <- roc(
  test_data %>% filter(race == "Black") %>% pull(status),
  test_predictions[test_data$race == "Black"]
)

# Output AUC
```

```r
auc_logistic_white <- auc(roc_logistic_white)
auc_logistic_black <- auc(roc_logistic_black)

print(paste("Logistic Test AUC for White:", auc_logistic_white))
```

Step_4: Model Fairness Evaluation

```
## [1] "Logistic Test AUC for White: 0.99040404040404"
```

```r
print(paste("Logistic Test AUC for Black:", auc_logistic_black))
```

```
## [1] "Logistic Test AUC for Black: 0.982142857142857"
```

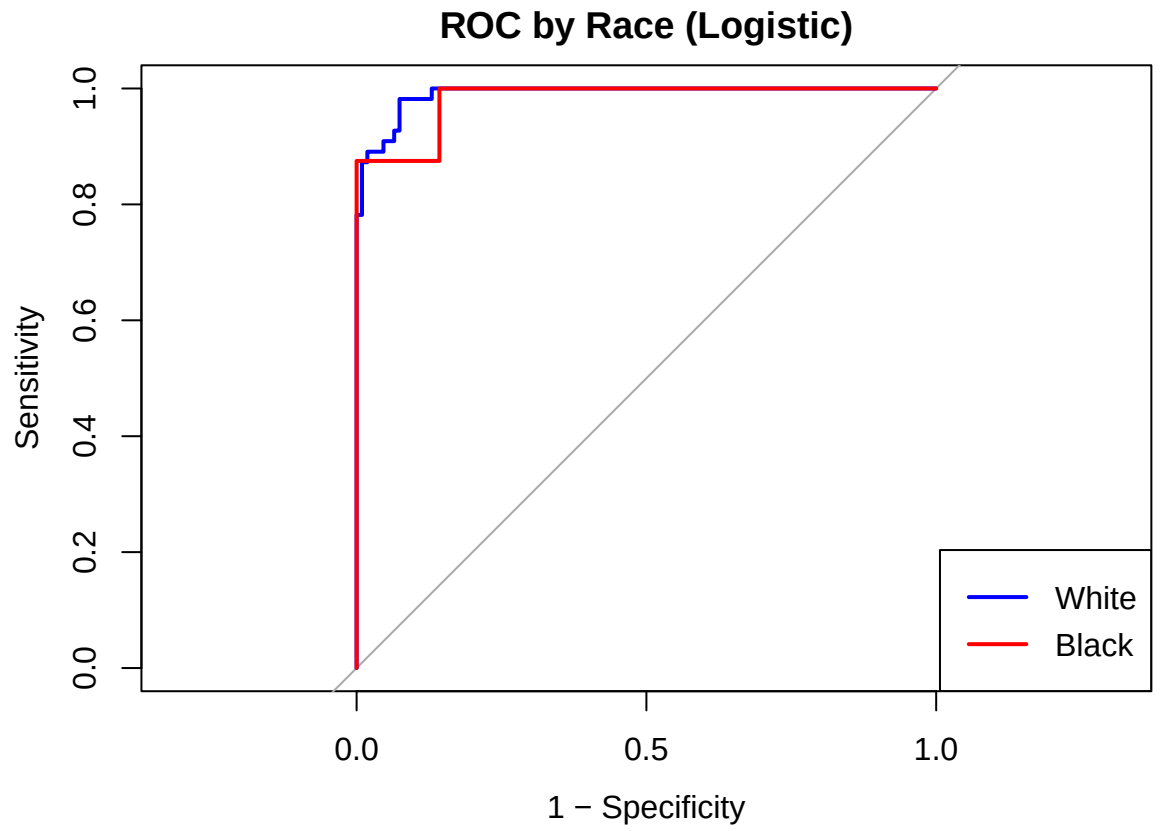
```r
# Cox model - Compute AUC by race
cox_risk_white <- predict(
  cox_model,
  newdata = cox_data %>% filter(race == "White"),
  type = "risk"
)
cox_risk_black <- predict(
  cox_model,
  newdata = cox_data %>% filter(race == "Black"),
  type = "risk"
)

# ROC curve - White group
roc_cox_white <- roc(
  cox_data %>% filter(race == "White") %>% pull(status),
  cox_risk_white
)

# ROC curve - Black group
roc_cox_black <- roc(
  cox_data %>% filter(race == "Black") %>% pull(status),
  cox_risk_black
)

# Output AUC
auc_cox_white <- auc(roc_cox_white)
auc_cox_black <- auc(roc_cox_black)

print(paste("Cox Model AUC for White:", auc_cox_white))
```

```
## [1] "Cox Model AUC for White: 0.971583791766396"
```

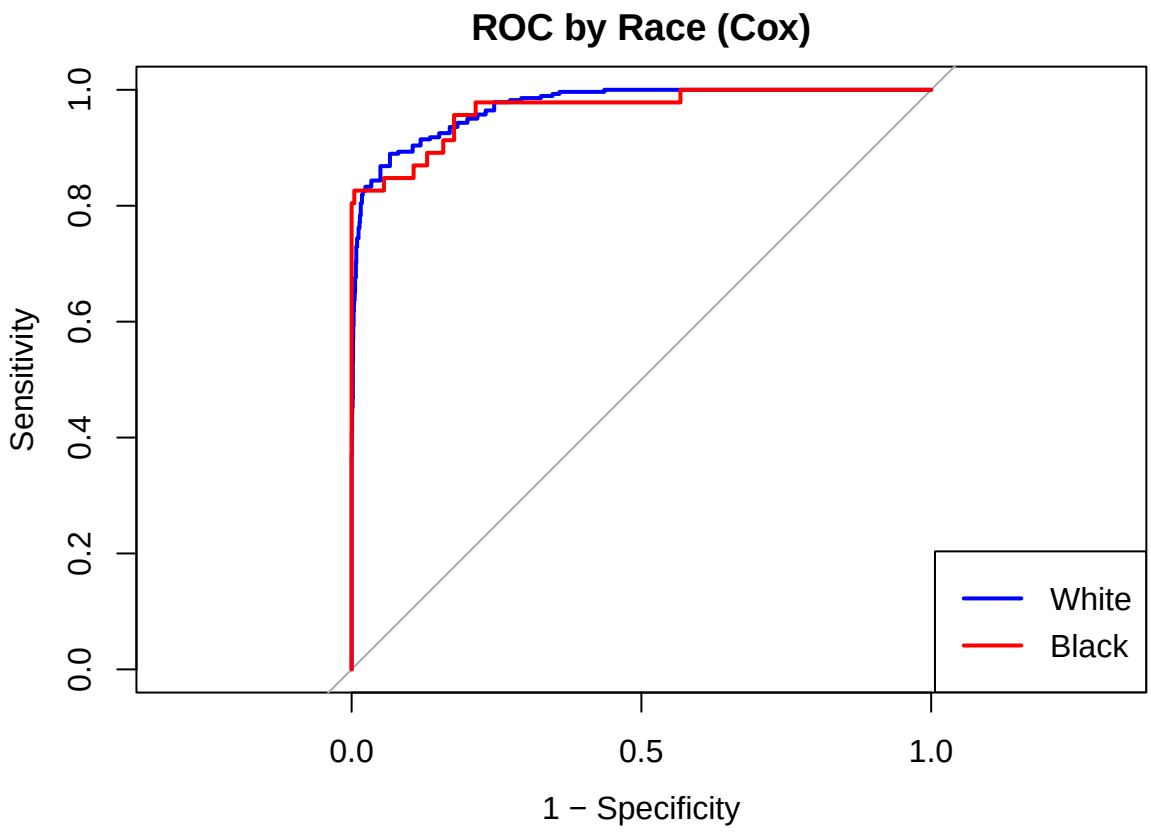
```r
print(paste("Cox Model AUC for Black:", auc_cox_black))
```

```
## [1] "Cox Model AUC for Black: 0.965419615773507"
```

```r
# Plot ROC curves by race for Logistic model
plot(roc_logistic_white, col = "blue", legacy.axes = TRUE, main = "ROC by Race (Logistic)")
lines(roc_logistic_black, col = "red")
legend("bottomright", legend = c("White", "Black"), col = c("blue", "red"), lwd = 2)
```




```r
# Plot ROC curves by race for Cox model
plot(roc_cox_white, col = "blue", legacy.axes = TRUE, main = "ROC by Race (Cox)")
lines(roc_cox_black, col = "red")
legend("bottomright", legend = c("White", "Black"), col = c("blue", "red"), lwd = 2)
```



Comments_1: - The figures demonstrate the predictive performance of Logistic and Cox models across racial groups (White and Black). For the Logistic model, the AUC is 0.9904 for the White group and 0.9821 for the Black group. For the Cox model, the AUC is 0.9716 for the White group and 0.9654 for the Black group. In the ROC curves, the White group's curve slightly outperforms the Black group's curve, indicating a minor performance difference.

Comments_2: - Both the Logistic and Cox models exhibit high predictive performance across racial groups, with slightly better results for the White group compared to the Black group, indicating minor unfairness. Although the differences are small, the potential bias against minority groups should be addressed, and further optimization is recommended to enhance fairness.

```
# Convert sparse matrix to a regular matrix
lasso_importance <- as.matrix(abs(coef_lasso))

# Convert matrix to a data frame
lasso_importance <- as.data.frame(lasso_importance)

# Add variable names as a column
lasso_importance$Variable <- rownames(lasso_importance)

# Rename columns for clarity
colnames(lasso_importance) <- c("Importance", "Variable")

# Reorder by importance
```

```r
lasso_importance <- lasso_importance[order(-lasso_importance$Importance), ]

# Display the feature importance table
knitr::kable(lasso_importance, caption = "Feature Importance for Logistic Model")
```

Step_5: Model Feature Importance Evaluation

Table 4: Feature Importance for Logistic Model

	Importance	Variable
survival_months_boxcox	4.2650954	survival_months_boxcox
differentiateUndifferentiated	3.3406388	differentiateUndifferentiated
raceOther	1.7033045	raceOther
regional_node_positive_log	1.2289846	regional_node_positive_log
(Intercept)	1.1126229	(Intercept)
t_stageT2	0.9403077	t_stageT2
progesterone_statusPositive	0.6840958	progesterone_statusPositive
differentiatePoorly differentiated	0.6297549	differentiatePoorly differentiated
raceWhite	0.5564617	raceWhite
status1:tumor_size_log	0.5409229	status1:tumor_size_log
regional_node_examined_log	0.2127777	regional_node_examined_log
age_scaled	0.1765491	age_scaled
estrogen_statusPositive	0.1671107	estrogen_statusPositive
age_scaled:t_stageT4	0.1264922	age_scaled:t_stageT4
differentiateWell differentiated	0.1250715	differentiateWell differentiated
marital_statusSingle	0.0955304	marital_statusSingle
n_stageN3	0.0913812	n_stageN3
stage_6thIIIC	0.0514802	stage_6thIIIC
grade3	0.0239706	grade3
gradeanaplastic; Grade IV	0.0118946	gradeanaplastic; Grade IV
marital_statusWidowed	0.0098700	marital_statusWidowed
tumor_size_log	0.0000000	tumor_size_log
marital_statusMarried	0.0000000	marital_statusMarried
marital_statusSeparated	0.0000000	marital_statusSeparated
t_stageT3	0.0000000	t_stageT3
t_stageT4	0.0000000	t_stageT4
n_stageN2	0.0000000	n_stageN2
stage_6thIIB	0.0000000	stage_6thIIB
stage_6thIIIA	0.0000000	stage_6thIIIA
stage_6thIIIB	0.0000000	stage_6thIIIB
grade2	0.0000000	grade2
a_stageRegional	0.0000000	a_stageRegional
age_scaled:t_stageT2	0.0000000	age_scaled:t_stageT2
age_scaled:t_stageT3	0.0000000	age_scaled:t_stageT3
survival_months_boxcox:progesterone_statusPositive	0.0000000	survival_months_boxcox:progesterone_statusPositive

Step_6: Model Calibration and Diagnostics Part_1: COX

```r
# Step 1: Check numeric variables for zero variance
print("Checking numeric variables for zero variance:")
```

```
## [1] "Checking numeric variables for zero variance:"
```

```r
zero_variance_vars <- apply(cox_data, 2, function(x) if (is.numeric(x)) var(x, na.rm = TRUE) == 0)
print("Variables with zero variance:")
```

```
## [1] "Variables with zero variance:"
```

```r
print(names(zero_variance_vars[zero_variance_vars]))
```

```
## NULL
```

```r
# Step 2: Check categorical variables for sparse levels
print("Checking categorical variables for sparse levels:")
```

```
## [1] "Checking categorical variables for sparse levels:"
```

```r
sparse_levels <- lapply(cox_data, function(x) if (is.factor(x)) table(x))
print("Sparse levels in categorical variables:")
```

```
## [1] "Sparse levels in categorical variables:"
```

```r
print(sparse_levels)
```

```
## $age
## NULL
##
## $race
## x
## Black Other White
##   261   293  3139
##
## $marital_status
## x
##  Divorced   Married Separated    Single   Widowed
##       443      2437        39       560       214
##
## $t_stage
## x
##   T1   T2   T3   T4
## 1493 1642  475   83
##
## $n_stage
## x
##   N1   N2   N3
## 2532  749  412
##
## $stage_6th
## x
##  IIA  IIB IIIA IIIB IIIC
## 1228 1035  962   56  412
##
## $differentiate
```

```
## x
## Moderately differentiated     Poorly differentiated          Undifferentiated
##                       2170                       993                        17
##       Well differentiated
##                        513
##
## $grade
## x
##                    1                    2                    3
##                  513                 2170                  993
## anaplastic; Grade IV
##                   17
##
## $a_stage
## x
##  Distant Regional
##       81     3612
##
## $tumor_size
## NULL
##
## $estrogen_status
## x
## Negative Positive
##      237     3456
##
## $progesterone_status
## x
## Negative Positive
##      624     3069
##
## $regional_node_examined
## NULL
##
## $regional_node_positive
## NULL
##
## $survival_months
## NULL
##
## $status
## NULL
##
## $tumor_size_log
## NULL
##
## $regional_node_examined_log
## NULL
##
## $regional_node_positive_log
## NULL
##
## $survival_months_log
## NULL
```

```
##
## $age_scaled
## NULL
##
## $survival_months_boxcox
## NULL
##
## $leverage
## NULL
##
## $std_residuals
## NULL
##
## $cooks_distance
## NULL
```

```
# Step 3: Perform individual Schoenfeld residual diagnostics to identify problematic variables
print("Performing individual Schoenfeld residual diagnostics:")
```

```
## [1] "Performing individual Schoenfeld residual diagnostics:"
```

```
problematic_vars <- c()
for (var in colnames(cox_model$x)) {
  tryCatch({
    single_var_test <- cox.zph(cox_model, transform = var)
    print(paste("Schoenfeld test for variable:", var))
    print(single_var_test)
  }, error = function(e) {
    print(paste("Error for variable:", var, "-", e$message))
    problematic_vars <<- c(problematic_vars, var)  # Collect problematic variables
  })
}
print("Problematic variables identified in Schoenfeld residual diagnostics:")
```

```
## [1] "Problematic variables identified in Schoenfeld residual diagnostics:"
```

```
print(problematic_vars)
```

```
## NULL
```

```
# Step 4: Adjust the Schoenfeld residual diagnostics for all variables
# Option 1: Use a different transformation (e.g., rank transformation)
print("Performing Schoenfeld residual diagnostics with rank transformation:")
```

```
## [1] "Performing Schoenfeld residual diagnostics with rank transformation:"
```

```
tryCatch({
  schoenfeld_test_rank <- cox.zph(cox_model, transform = "rank")
  print("Schoenfeld test results with rank transformation:")
  print(schoenfeld_test_rank)
```

```
  plot(schoenfeld_test_rank, main = "Schoenfeld Residuals with Rank Transformation")
}, error = function(e) {
  print(paste("Error in Schoenfeld test with rank transformation:", e$message))
})
```

```
## [1] "Error in Schoenfeld test with rank transformation: system is computationally singular: reciproca
```

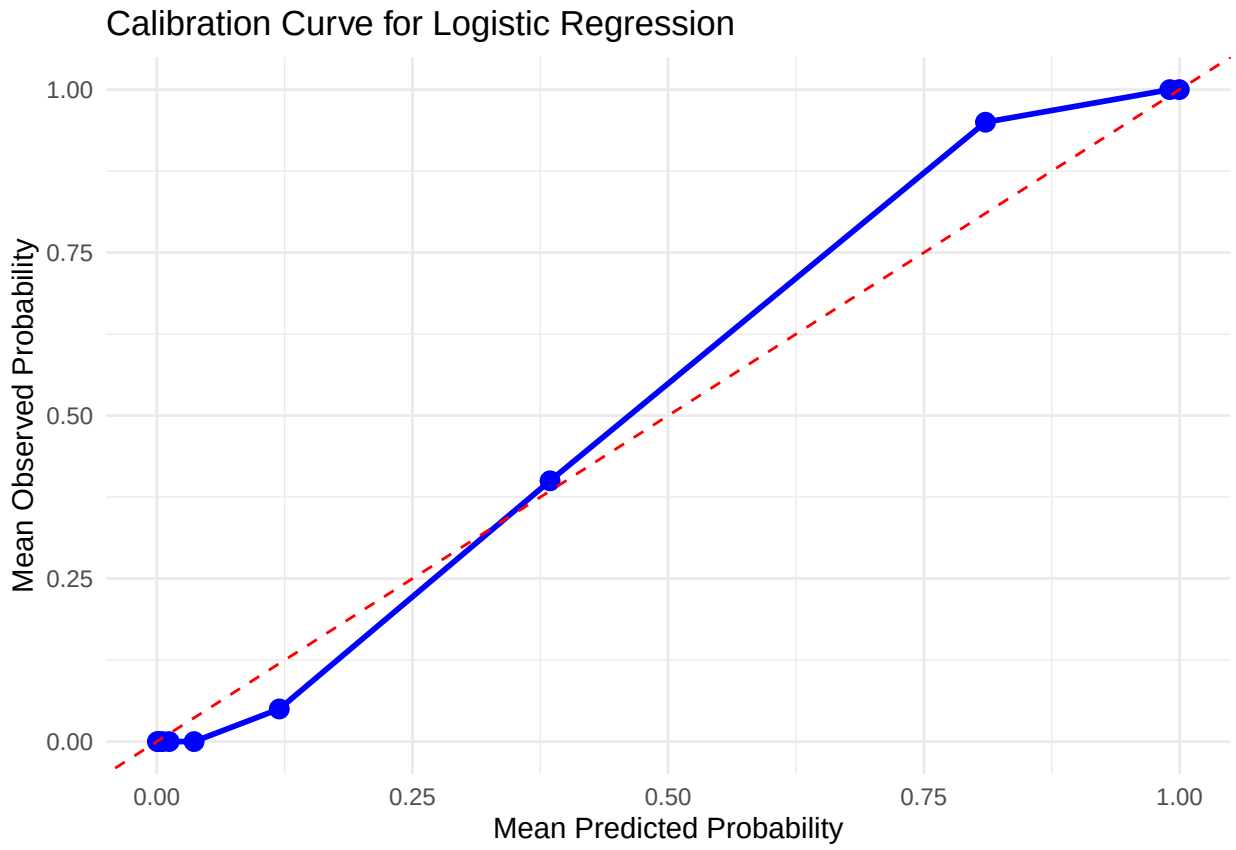
```
# Step 5: If specific problematic variables are identified, exclude them from diagnostics
if (length(problematic_vars) > 0) {
  print("Excluding problematic variables from diagnostics:")
  reduced_cox_formula <- as.formula(paste("Surv(survival_months_boxcox, status) ~ ",
                                          paste(setdiff(colnames(cox_model$x), problematic_vars), collap
  reduced_cox_model <- coxph(reduced_cox_formula, data = cox_data)

  # Perform Schoenfeld residual diagnostics on reduced model
  print("Performing Schoenfeld residual diagnostics on reduced model:")
  tryCatch({
    schoenfeld_test_reduced <- cox.zph(reduced_cox_model)
    print("Schoenfeld test results for reduced model:")
    print(schoenfeld_test_reduced)
    plot(schoenfeld_test_reduced, main = "Schoenfeld Residuals for Reduced Model")
  }, error = function(e) {
    print(paste("Error in Schoenfeld test for reduced model:", e$message))
  })
} else {
  print("No problematic variables detected for exclusion.")
}
```

```
## [1] "No problematic variables detected for exclusion."
```

Part_2: Logistic Models

```
# Logistic Model Manual Calibration Curve (Fixed Version)

# Step 1: Ensure status is numeric (convert if necessary)
test_data <- test_data %>%
  mutate(status = as.numeric(as.character(status)))  # Convert status to numeric if not already

# Step 2: Add predicted probabilities and bin into deciles
test_data <- test_data %>%
  mutate(predicted_prob = test_predictions) %>%  # Add predicted probabilities
  mutate(decile = ntile(predicted_prob, 10))  # Bin into 10 deciles

# Step 3: Calculate mean predicted and observed probabilities for each decile
library(dplyr)
calibration_data <- test_data %>%
  group_by(decile) %>%
  summarise(
    mean_predicted = mean(predicted_prob, na.rm = TRUE),  # Average predicted probability
    mean_observed = mean(status, na.rm = TRUE)  # Average observed status (0/1)
  )
```

```r
# Step 4: Remove rows with NA (if any) from calibration data
calibration_data <- calibration_data %>%
  filter(!is.na(mean_predicted) & !is.na(mean_observed))

# Step 5: Plot calibration curve
library(ggplot2)
ggplot(calibration_data, aes(x = mean_predicted, y = mean_observed)) +
  geom_point(size = 3, color = "blue") +  # Points for observed vs predicted
  geom_line(color = "blue", size = 1) +   # Line connecting the points
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +  # Perfect calibration li
  labs(
    x = "Mean Predicted Probability",
    y = "Mean Observed Probability",
    title = "Calibration Curve for Logistic Regression"
  ) +
  theme_minimal()
```




```r
# Descriptive Summary Table
summary_stats <- data_clean_no_outliers %>%
  summarise(
    Age = list(summary(age)),
    TumorSize = list(summary(tumor_size_log)),
```

```
    SurvivalMonths = list(summary(survival_months_boxcox)),
    NodeExamined = list(summary(regional_node_examined_log))
  )

knitr::kable(summary_stats, caption = "Summary Statistics for Key Variables")
```

Step_7: Report Supplement and Results Summary

Table 5: Summary Statistics for Key Variables

Age	TumorSize	SurvivalMonths	NodeExamined
30.0000, 47.0000, 54.0000, 54.0677, 61.0000, 69.0000	0.6931472, 2.8332133, 3.2580965, 3.2482217, 3.6375862, 4.9487599	0.9333737, 12.3040417, 13.5665533, 13.2040930, 14.6847863, 15.5904935	0.6931472, 2.3025851, 2.7080502, 2.5623464, 2.9957323, 4.1271344

```
# Model Results Summary Table
result_summary <- data.frame(
  Model = c("Logistic Regression", "Cox Proportional Hazards"),
  Train_AUC = c(auc_train, NA),
  Test_AUC = c(auc_test, auc_cox),
  Significant_Features = c(nrow(selected_variables_df), nrow(significant_coefficients))
)

knitr::kable(result_summary, caption = "Summary of Model Results")
```

Table 6: Summary of Model Results

Model	Train_AUC	Test_AUC	Significant_Features
Logistic Regression	0.9803298	0.9908088	21
Cox Proportional Hazards	NA	0.9702371	21