# Data analysis

12/09/2023

## Libraries

```
library(tidyverse)
library(readr)
library(boot)
library(table1)
library(gridExtra)
library(MASS)
library(car)
library(dplyr)
library(leaps)
library(corrplot)
```

## Data Clean

```
breastcancer_data =
  read_csv("Project_2_data.csv") |>
  janitor::clean_names()
```

```
## Rows: 4024 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): Race, Marital Status, T Stage, N Stage, 6th Stage, differentiate, ...
## dbl  (5): Age, Tumor Size, Regional Node Examined, Reginol Node Positive, Su...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(breastcancer_data)
```

```
##       age            race           marital_status       t_stage
##  Min.   :30.00   Length:4024        Length:4024        Length:4024
##  1st Qu.:47.00   Class :character   Class :character   Class :character
##  Median :54.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :53.97
##  3rd Qu.:61.00
##  Max.   :69.00
##    n_stage           x6th_stage        differentiate         grade
##  Length:4024        Length:4024        Length:4024        Length:4024
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    a_stage            tumor_size     estrogen_status    progesterone_status
##  Length:4024        Min.   :  1.00   Length:4024        Length:4024
##  Class :character   1st Qu.: 16.00   Class :character   Class :character
##  Mode  :character   Median : 25.00   Mode  :character   Mode  :character
##                     Mean   : 30.47
##                     3rd Qu.: 38.00
##                     Max.   :140.00
##  regional_node_examined reginol_node_positive survival_months
##  Min.   : 1.00          Min.   : 1.000        Min.   :  1.0
##  1st Qu.: 9.00          1st Qu.: 1.000        1st Qu.: 56.0
##  Median :14.00          Median : 2.000        Median : 73.0
##  Mean   :14.36          Mean   : 4.158        Mean   : 71.3
##  3rd Qu.:19.00          3rd Qu.: 5.000        3rd Qu.: 90.0
##  Max.   :61.00          Max.   :46.000        Max.   :107.0
##     status
##  Length:4024
##  Class :character
##  Mode  :character
##
##
##
```

```
bc = breastcancer_data |>
  mutate(
    race=case_when(
      race == "White" ~ 1,
      race == "Black" ~ 2,
      race == "Other" ~ 3),
    marital_status=case_when(
      marital_status == "Married" ~ 1,
      marital_status == "Divorced" ~ 2,
      marital_status == "Single" ~ 3,
      marital_status == "Widowed" ~ 4,
      marital_status == "Separated" ~ 5),
    t_stage=case_when(
      t_stage == "T1" ~ 1,
      t_stage == "T2" ~ 2,
      t_stage == "T3" ~ 3,
      t_stage == "T4" ~ 4),
    n_stage=case_when(
      n_stage == "N1" ~ 1,
      n_stage == "N2" ~ 2,
      n_stage == "N3" ~ 3),
    x6th_stage=case_when(
      x6th_stage == "IIA" ~ 1,
      x6th_stage == "IIIA" ~ 2,
      x6th_stage == "IIIC" ~ 3,
      x6th_stage == "IIB" ~ 4,
      x6th_stage == "IIIB" ~ 5),
```

```
    differentiate=case_when(
      differentiate == "Poorly differentiated" ~ 1,
      differentiate == "Moderately differentiated" ~ 2,
      differentiate == "Well differentiated" ~ 3,
      differentiate == "Undifferentiated" ~ 4),
    grade=case_when(
      grade == "1" ~ 1,
      grade == "2" ~ 2,
      grade == "3" ~ 3,
      grade == "anaplastic; Grade IV" ~ 4),
    a_stage=case_when(
      a_stage == "Regional" ~ 1,
      a_stage == "Distant" ~ 0),
    estrogen_status=case_when(
      estrogen_status == "Positive" ~ 1,
      estrogen_status == "Negative" ~ 0),
    progesterone_status=case_when(
      progesterone_status == "Positive" ~ 1,
      progesterone_status == "Negative" ~ 0),
    status=case_when(
      status == "Alive" ~ 1,
      status == "Dead" ~ 0)
    )
```

### Descriptive statistics for all variables

```
summary(bc)
```

```
##       age             race       marital_status     t_stage
##  Min.   :30.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:47.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :54.00   Median :1.000   Median :1.000   Median :2.000
##  Mean   :53.97   Mean   :1.231   Mean   :1.646   Mean   :1.785
##  3rd Qu.:61.00   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :69.00   Max.   :3.000   Max.   :5.000   Max.   :4.000
##     n_stage        x6th_stage    differentiate       grade
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :1.000   Median :2.000   Median :2.000   Median :2.000
##  Mean   :1.438   Mean   :2.405   Mean   :1.868   Mean   :2.151
##  3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :5.000   Max.   :4.000   Max.   :4.000
##     a_stage         tumor_size     estrogen_status  progesterone_status
##  Min.   :0.0000   Min.   :  1.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.: 16.00   1st Qu.:1.0000   1st Qu.:1.0000
##  Median :1.0000   Median : 25.00   Median :1.0000   Median :1.0000
##  Mean   :0.9771   Mean   : 30.47   Mean   :0.9332   Mean   :0.8265
##  3rd Qu.:1.0000   3rd Qu.: 38.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :140.00   Max.   :1.0000   Max.   :1.0000
##  regional_node_examined reginol_node_positive survival_months     status
##  Min.   : 1.00          Min.   : 1.000        Min.   :  1.0   Min.   :0.0000
```

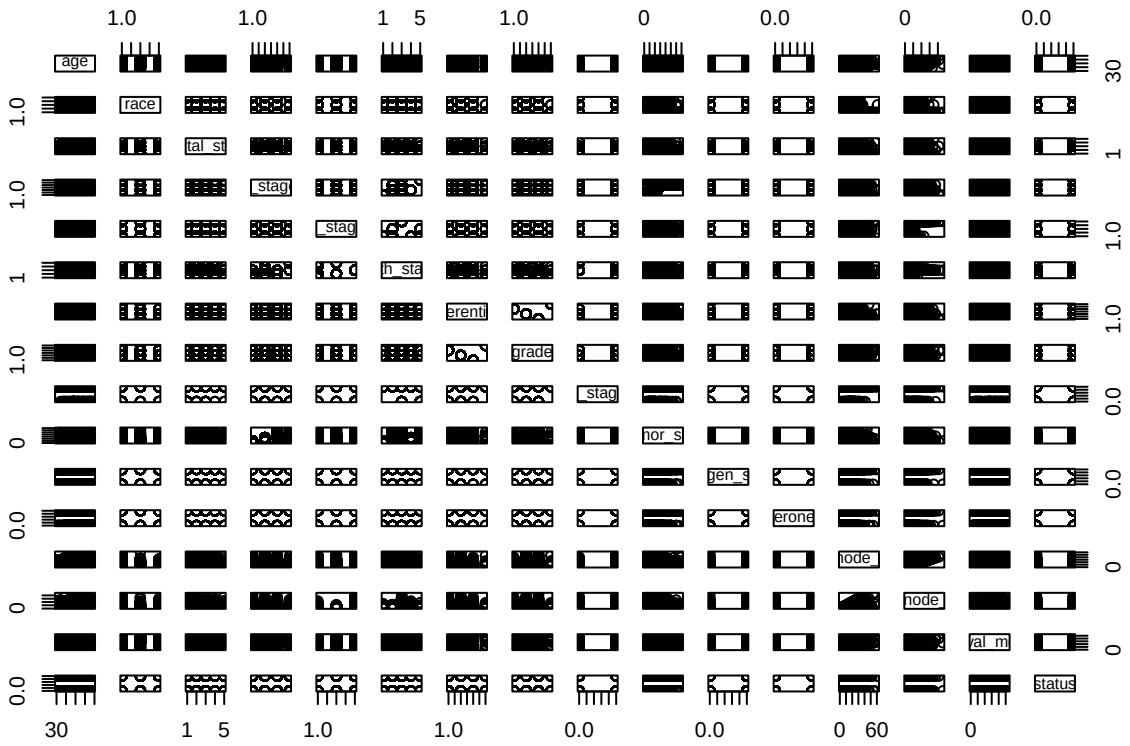
```
##  1st Qu.: 9.00          1st Qu.: 1.000        1st Qu.: 56.0   1st Qu.:1.0000
##  Median :14.00          Median : 2.000        Median : 73.0   Median :1.0000
##  Mean   :14.36          Mean   : 4.158        Mean   : 71.3   Mean   :0.8469
##  3rd Qu.:19.00          3rd Qu.: 5.000        3rd Qu.: 90.0   3rd Qu.:1.0000
##  Max.   :61.00          Max.   :46.000        Max.   :107.0   Max.   :1.0000
```

We change race, marital_status, t_stage, n_stage, x6th_stage, differentiate, and grade into multiple numeric levels, while a_stage, estrogen_status, progesterone_status, and status to binary levels. The above variables are categorical variables.

And age, tumor_size, regional_node_examined, reginol_node_positive, and survival_months are numeric variables.

## Covariance and Correlation

```
plot(bc)
```



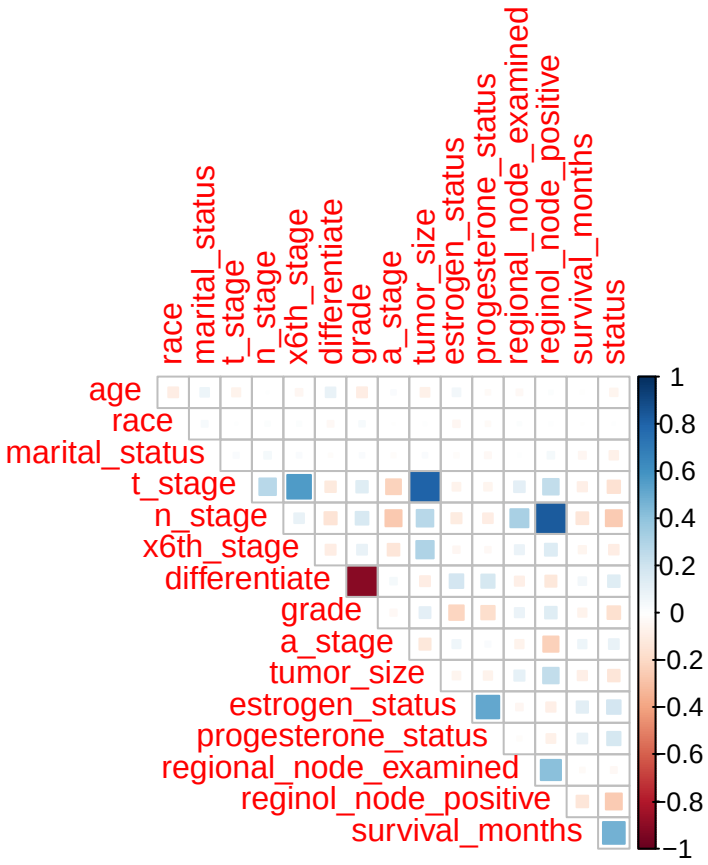

```
cor(bc) |>
  knitr::kable(digits=4,caption="Correlation for all variables")
```

Table 1: Correlation for all variables

| | age | race | marital_status | t_stage | n_stage | x6th_stage | differentiate | grade | a_stage | tumor_size | estrogen_status | progesterone_status | regional_node_examined | reginol_node_positive | survival_months | status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age | 1.0000 | -0.0970 | 0.0755 | -0.0669 | 0.0029 | -0.0450 | -0.0932 | -0.0993 | -0.0209 | -0.0772 | -0.0598 | -0.0213 | -0.0333 | -0.0126 | -0.0094 | -0.0559 |
| race | -0.0970 | 1.0000 | 0.0349 | 0.0036 | 0.0191 | 0.0173 | -0.0336 | -0.0301 | -0.0020 | -0.0071 | -0.0425 | -0.0209 | -0.0113 | -0.0090 | -0.0025 | 0.0042 |
| marital_status | 0.0755 | 0.0349 | 1.0000 | 0.0257 | 0.0457 | 0.0221 | -0.0117 | -0.0206 | -0.0174 | -0.0207 | -0.0198 | -0.0357 | -0.0077 | -0.0443 | -0.0487 | -0.0731 |
| t_stage | -0.0669 | 0.0036 | 0.0257 | 1.0000 | 0.2770 | 0.5637 | -0.1102 | -0.1315 | -0.2211 | -0.8092 | -0.0610 | -0.0576 | -0.1141 | -0.2431 | -0.0857 | -0.1547 |
| n_stage | 0.0029 | 0.0191 | 0.0457 | 0.2770 | 1.0000 | 0.0939 | -0.1488 | -0.1625 | -0.2606 | -0.2779 | -0.1020 | -0.0937 | -0.3283 | -0.8381 | -0.1396 | -0.2558 |
| x6th_stage | 0.0450 | 0.0173 | 0.0221 | 0.5637 | 0.0939 | 1.0000 | -0.0999 | -0.0972 | -0.1372 | -0.3034 | -0.0417 | -0.0309 | -0.0826 | -0.1427 | -0.0536 | -0.0919 |
| differentiate | 0.0932 | -0.0336 | -0.0117 | -0.1102 | -0.1488 | -0.0999 | 1.0000 | -0.9083 | -0.0395 | -0.1868 | -0.2113 | -0.1799 | -0.0834 | -0.1229 | -0.0677 | -0.1614 |
| grade | -0.0993 | 0.0301 | 0.0206 | 0.1315 | 0.1625 | 0.0972 | -0.9083 | 1.0000 | -0.1239 | 0.1194 | -0.0596 | -0.0699 | -0.0690 | -0.2328 | -0.0869 | -0.1342 |
| a_stage | 0.0209 | -0.0020 | -0.0174 | -0.2211 | -0.2606 | -0.1372 | 0.0395 | -0.1239 | 1.0000 | -0.0656 | -0.0265 | -0.0699 | -0.1044 | -0.2423 | -0.1285 | -0.1847 |
| tumor_size | 0.0772 | 0.0071 | 0.0207 | 0.8092 | 0.2779 | 0.3034 | -0.0995 | 0.1194 | -0.1239 | 1.0000 | -0.0596 | -0.0699 | -0.0448 | -0.0860 | -0.0960 | -0.1771 |
| estrogen_status | 0.0598 | 0.0425 | 0.0198 | 0.0610 | 0.1020 | 0.0417 | -0.2113 | 0.0596 | -0.0265 | 0.0596 | 1.0000 | 0.5133 | -0.0181 | -0.0781 | -0.0221 | -0.0348 |
| progesterone_status | 0.0213 | 0.0209 | 0.0357 | 0.0576 | 0.0937 | 0.0309 | -0.1799 | 0.0699 | -0.0699 | 0.0699 | 0.5133 | 1.0000 | -0.0181 | -0.0781 | -0.1352 | -0.2566 |
| regional_node_examined | 0.0333 | 0.0113 | 0.0077 | 0.1141 | 0.3283 | 0.0826 | -0.0834 | 0.0844 | -0.0690 | 0.1044 | -0.0448 | -0.0181 | 1.0000 | 0.4116 | -0.0221 | -0.0348 |
| reginol_node_positive | 0.0126 | 0.0090 | 0.0443 | 0.2431 | 0.8381 | 0.1427 | -0.1229 | 0.1353 | -0.2328 | 0.2423 | -0.0860 | -0.0781 | -0.4116 | 1.0000 | -0.1352 | -0.2566 |
| survival_months | 0.0094 | 0.0025 | 0.0487 | 0.0857 | 0.1396 | 0.0536 | -0.0677 | 0.0869 | -0.1285 | 0.0960 | -0.0221 | -0.1352 | -0.0221 | -0.1352 | 1.0000 | 0.4765 |
| status | -0.0559 | 0.0042 | 0.0731 | 0.1547 | 0.2558 | 0.0919 | -0.1614 | 0.1342 | -0.1847 | 0.1771 | -0.0348 | -0.2566 | -0.0348 | -0.2566 | 0.4765 | 1.0000 |

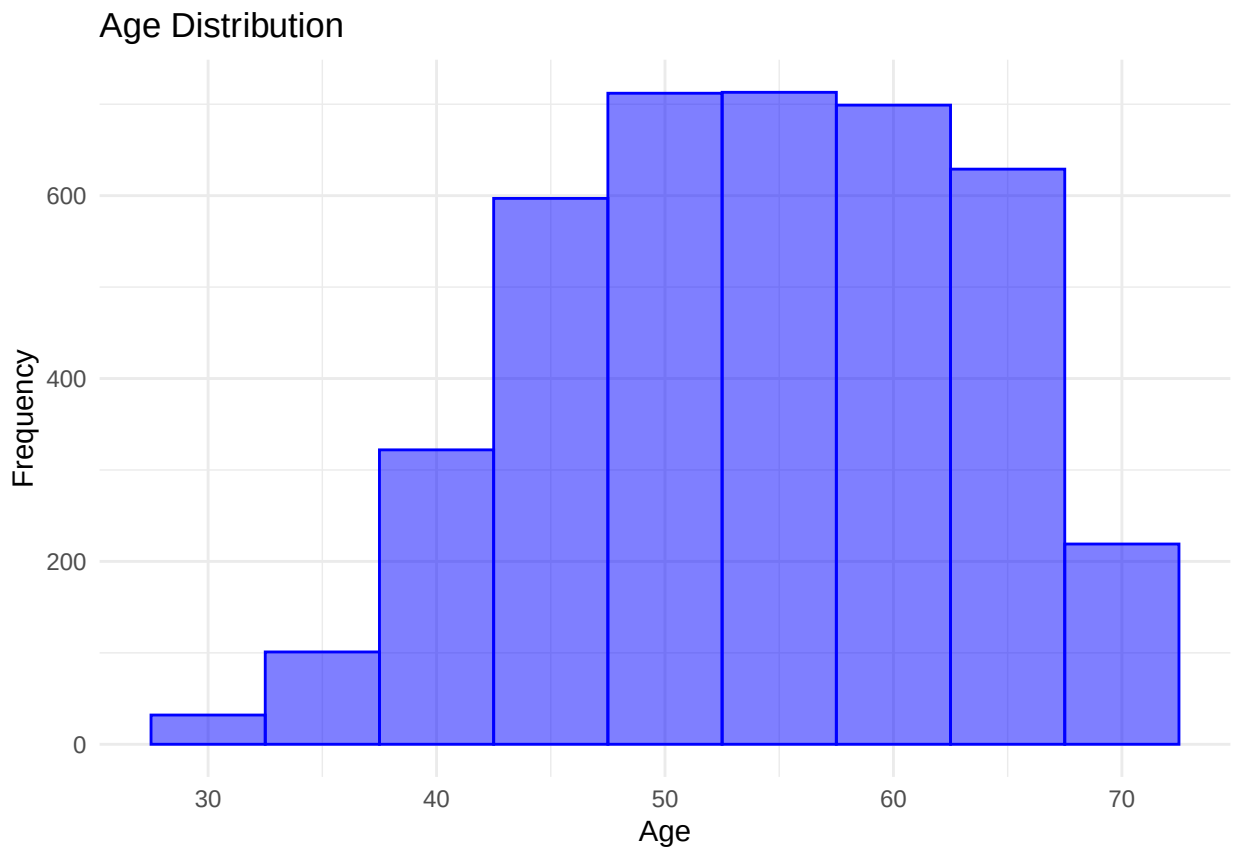
**Another plot for correlation**

```
corrplot(cor(bc), method = "square", type = "upper", diag = FALSE)
```

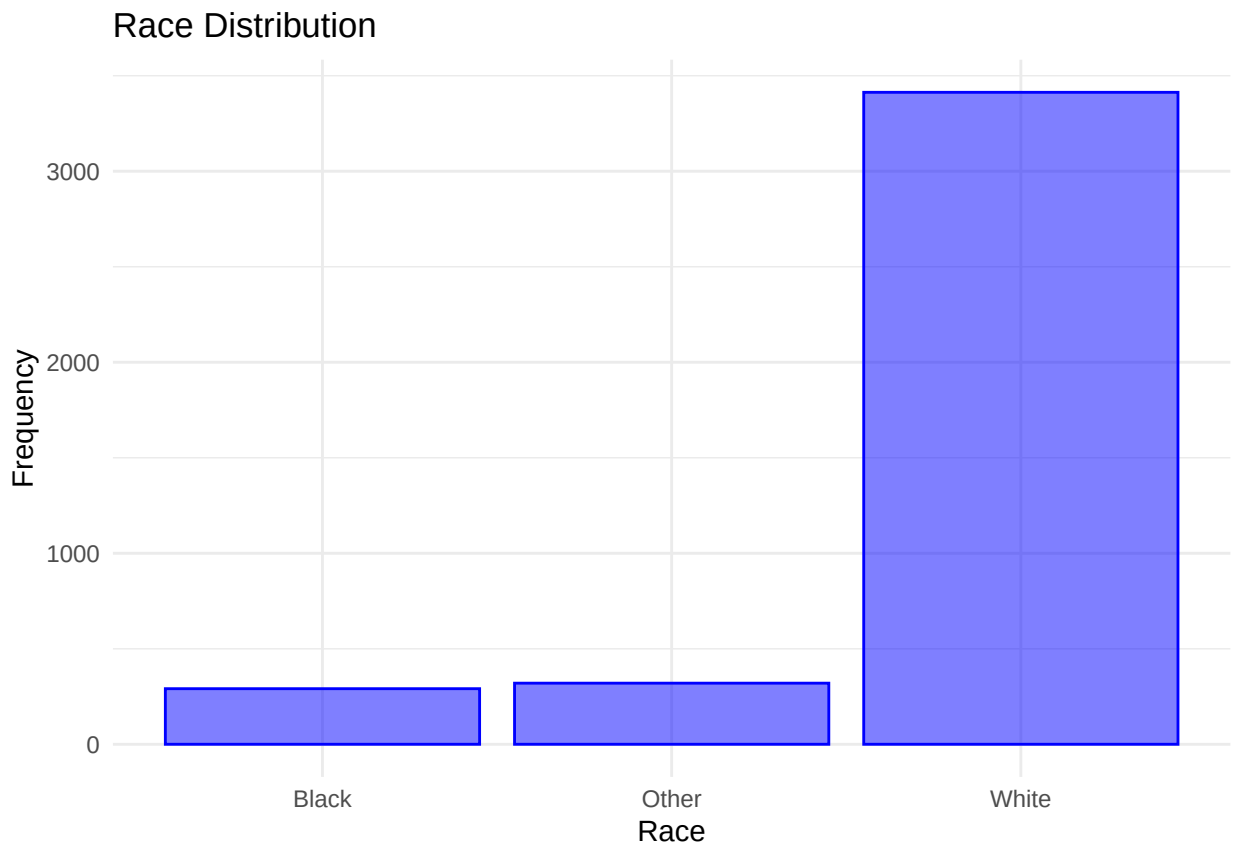



### Exploratory visualisation

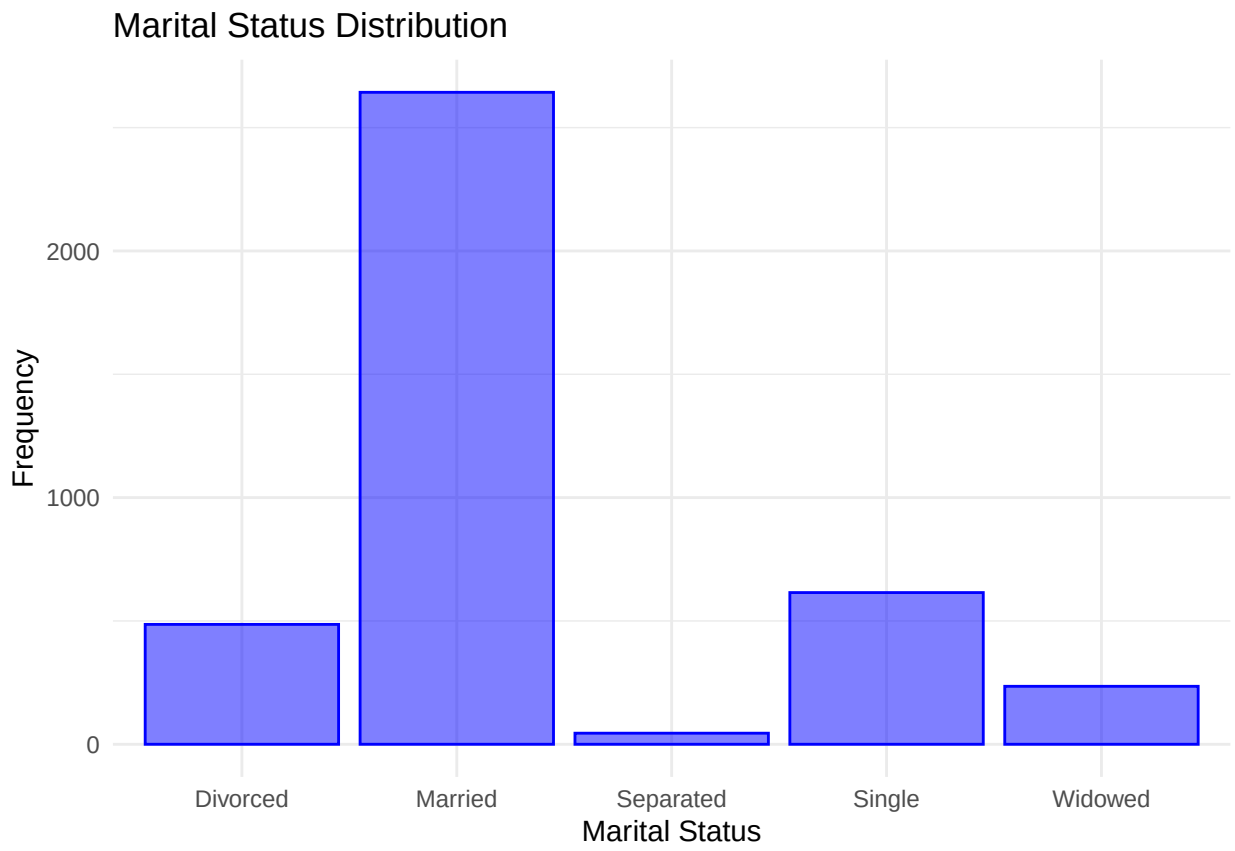
```
plot1age =
  breastcancer_data|>
  ggplot(aes(x = age)) +
  geom_histogram(color = "blue", fill = alpha("blue", 0.5), binwidth = 5) +
  theme_minimal() +
  labs(
    title = "Age Distribution",
    x = "Age",
    y = "Frequency"
  )
plot1age
```



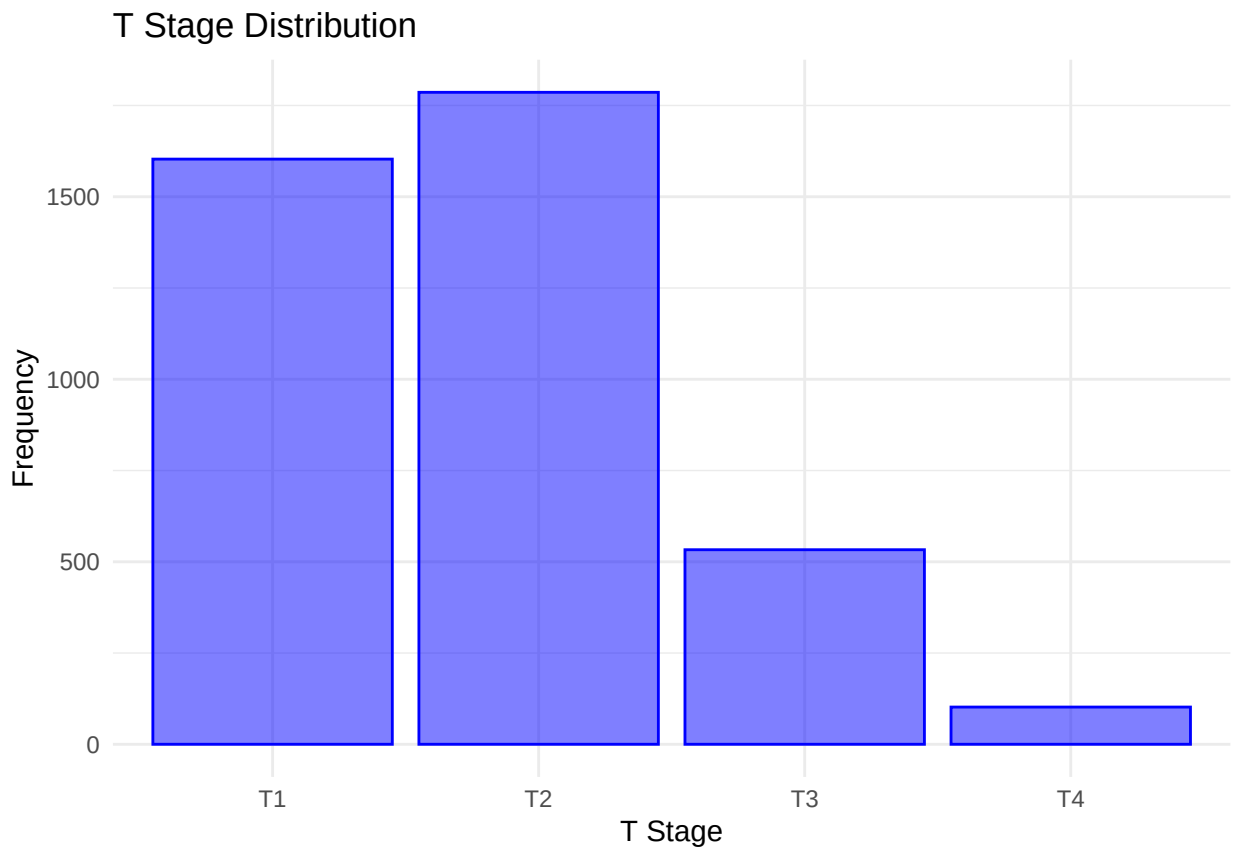

```
plot2race =
  breastcancer_data|>
  ggplot(aes(x = race)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Race Distribution",
    x = "Race",
    y = "Frequency"
  )
plot2race
```



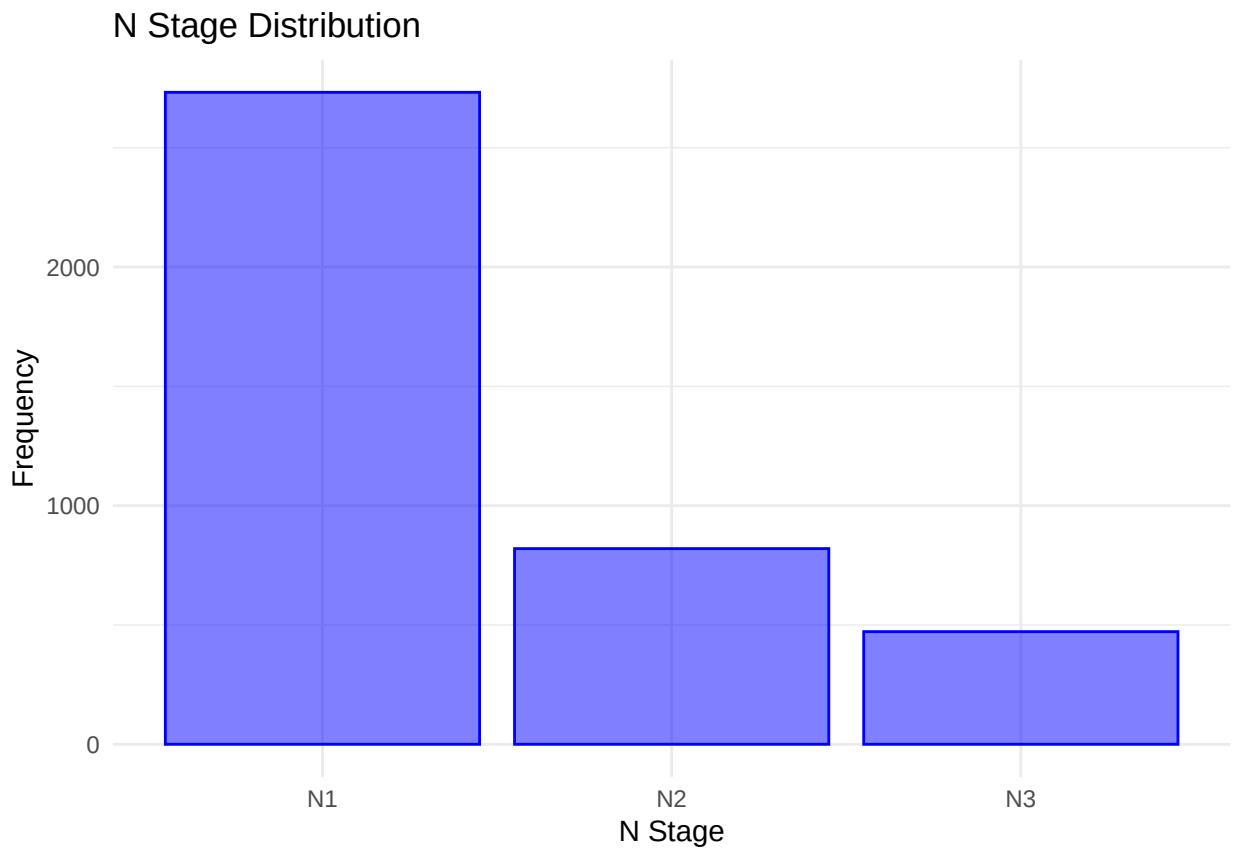

```
plot3marital =
  breastcancer_data|>
  ggplot(aes(x = marital_status)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Marital Status Distribution",
    x = "Marital Status",
    y = "Frequency"
  )

plot3marital
```



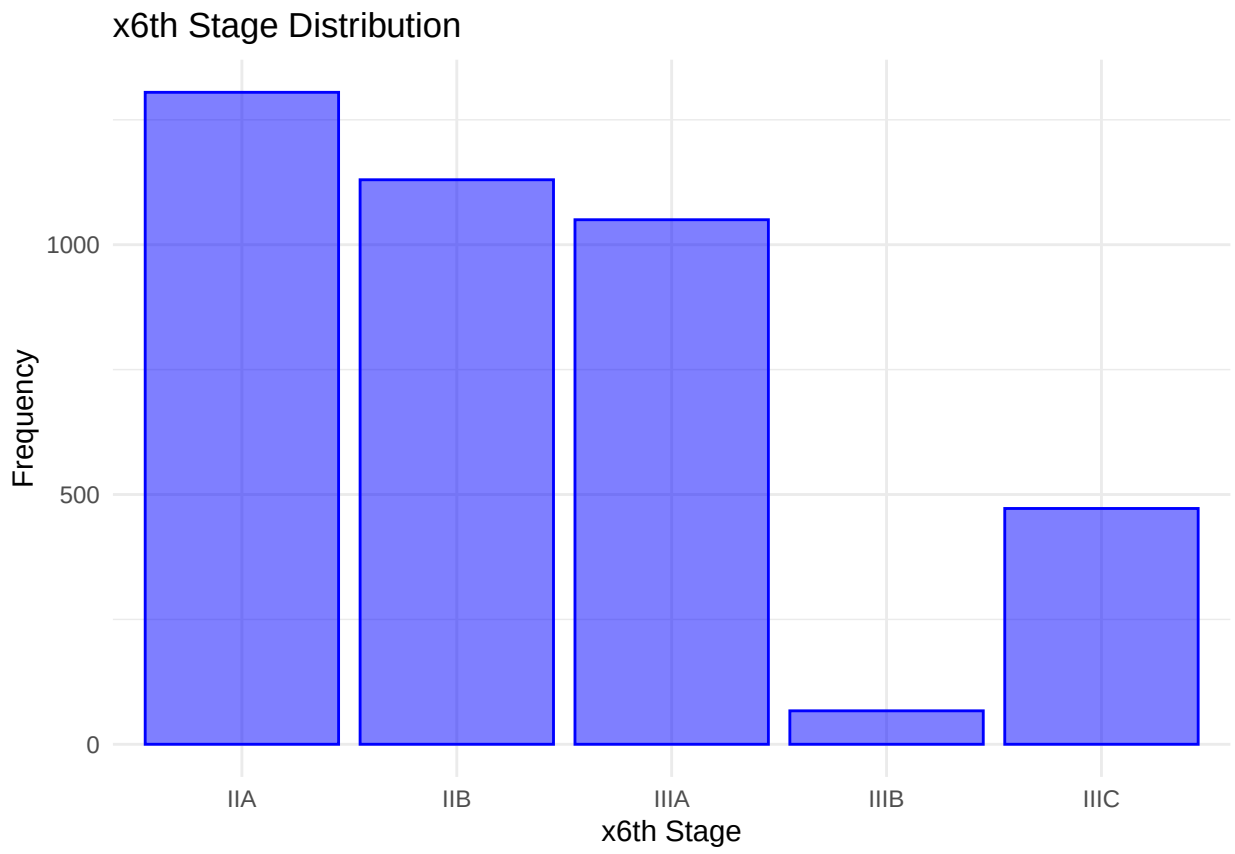

```
plot4tstage =
  breastcancer_data|>
  ggplot(aes(x = t_stage)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "T Stage Distribution",
    x = "T Stage",
    y = "Frequency"
  )

plot4tstage
```



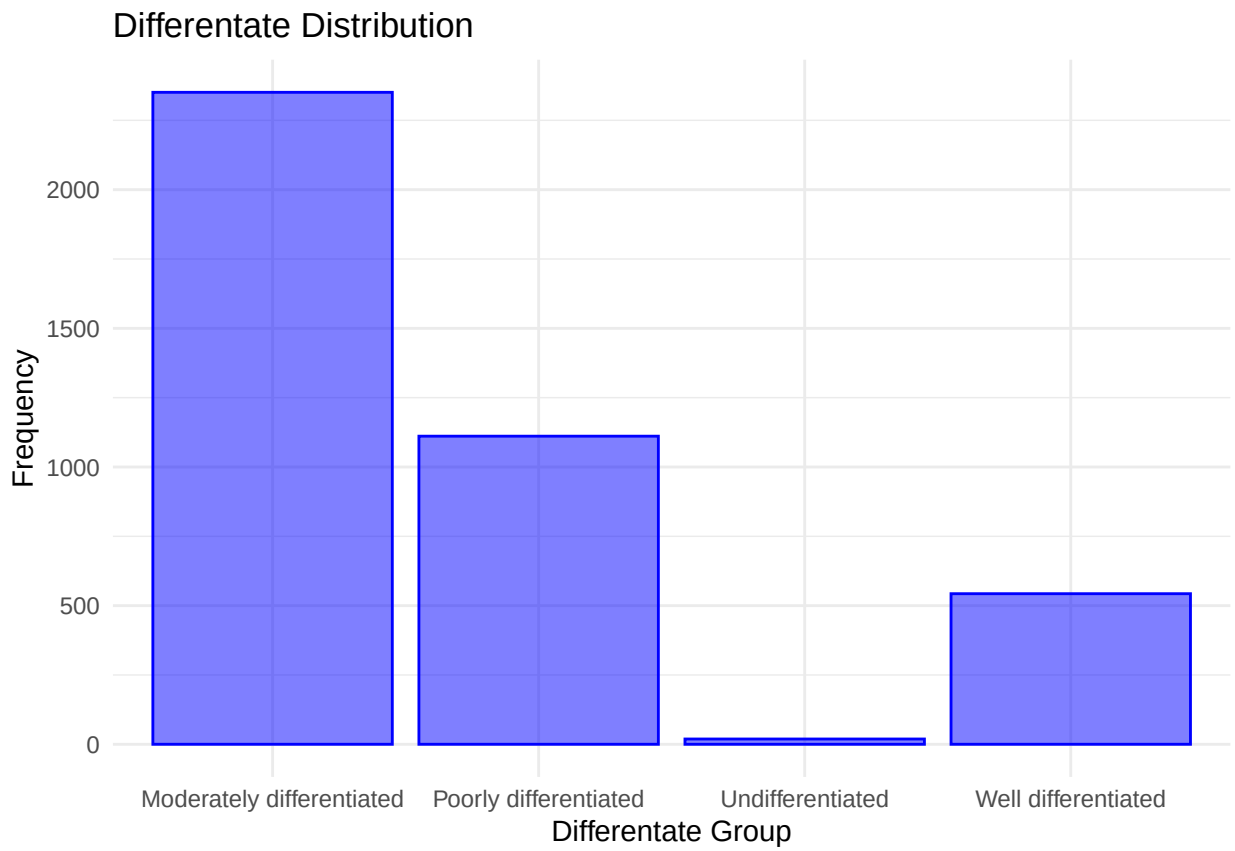

```
plot5nstage =
  breastcancer_data|>
  ggplot(aes(x = n_stage)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "N Stage Distribution",
    x = "N Stage",
    y = "Frequency"
  )

plot5nstage
```



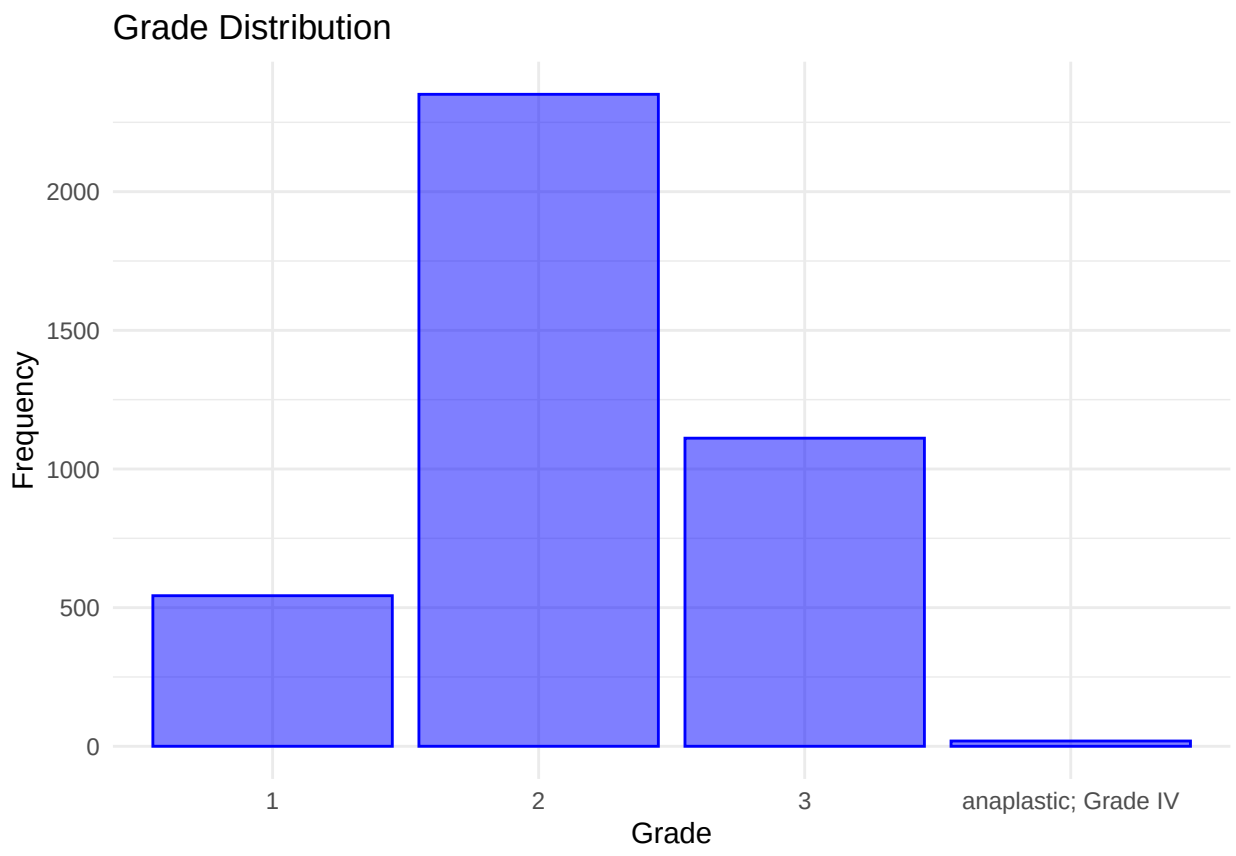

```
plot6x6thstage =
  breastcancer_data|>
  ggplot(aes(x = x6th_stage)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "x6th Stage Distribution",
    x = "x6th Stage",
    y = "Frequency"
  )

plot6x6thstage
```



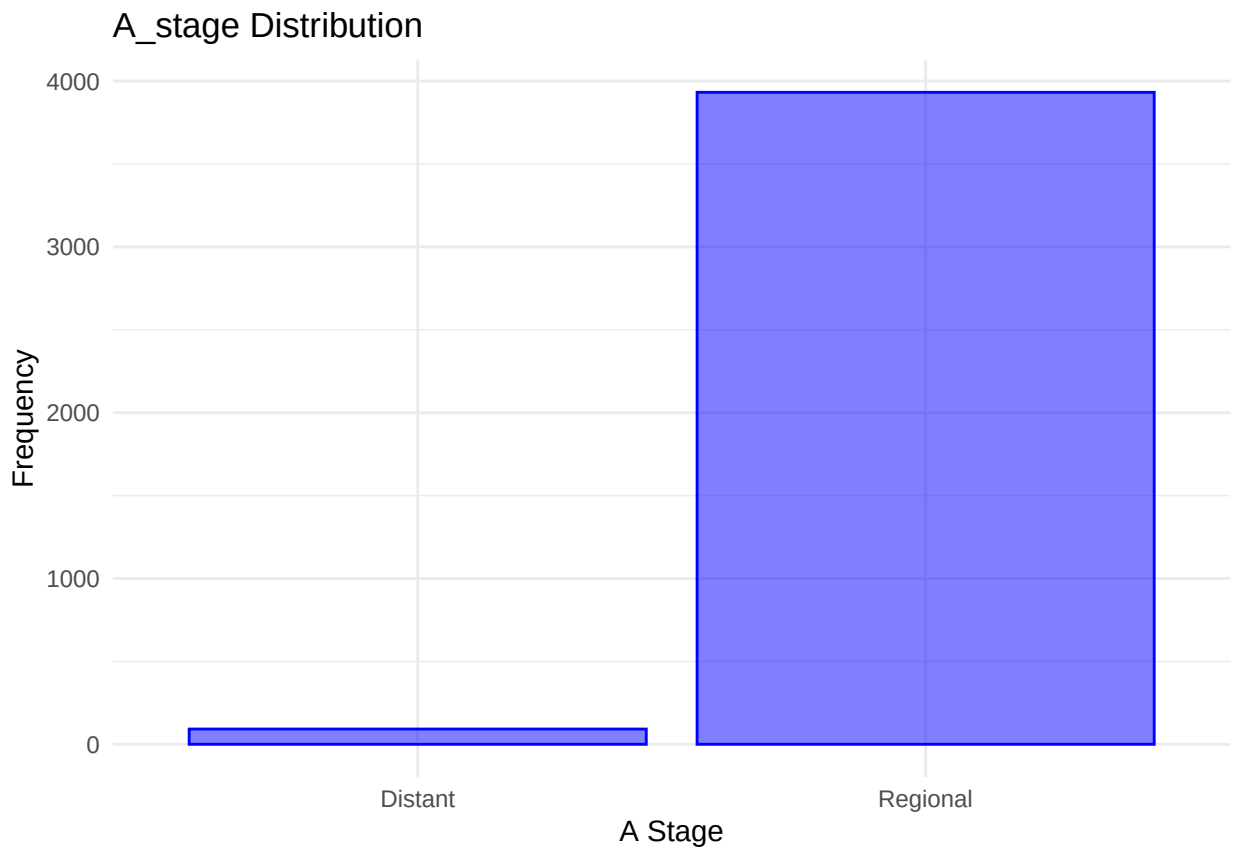

```
plot7differentate =
  breastcancer_data|>
  ggplot(aes(x = differentiate)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Differentate Distribution",
    x = "Differentate Group",
    y = "Frequency"
  )

plot7differentate
```




```
plot8grade =
  breastcancer_data|>
  ggplot(aes(x = grade)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Grade Distribution",
    x = "Grade",
    y = "Frequency"
  )

plot8grade
```

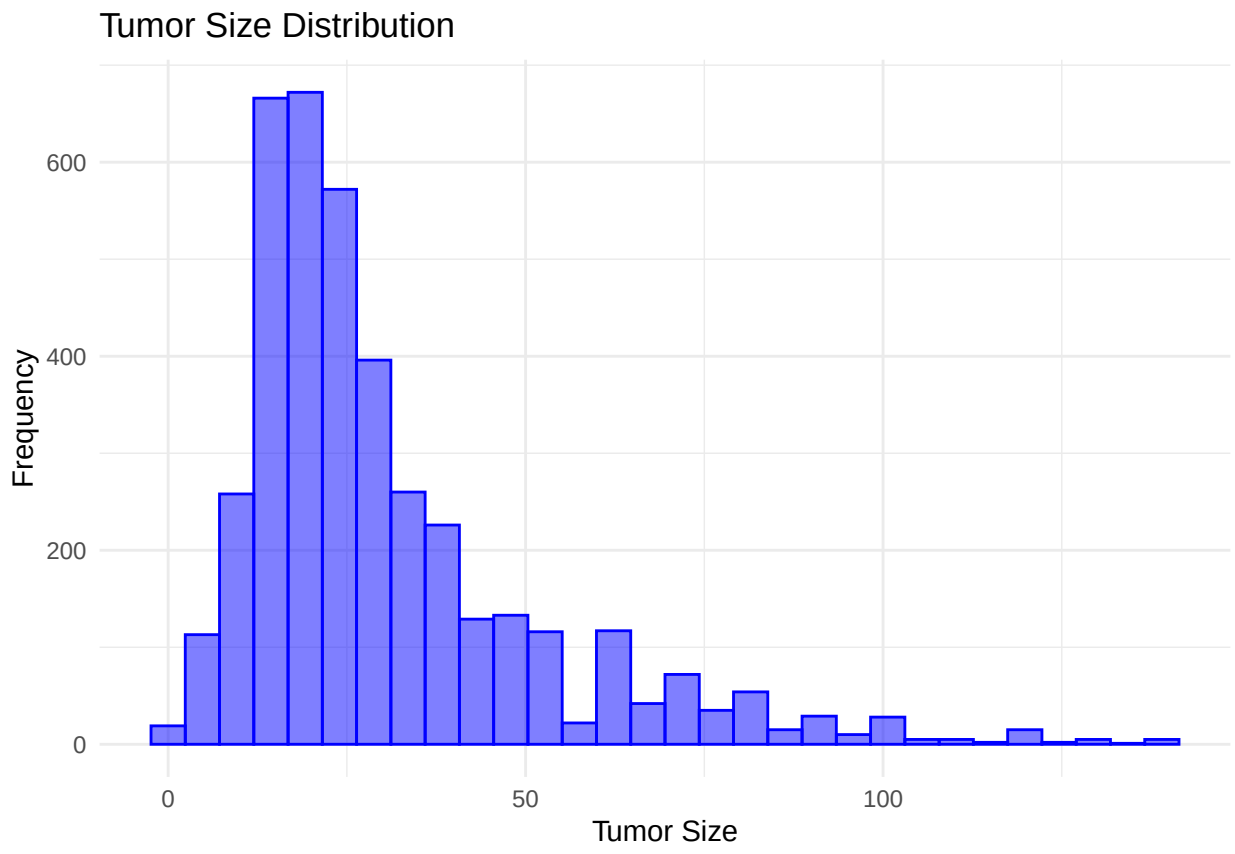



```
plot9astage =
  breastcancer_data|>
  ggplot(aes(x = a_stage)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "A_stage Distribution",
    x = "A Stage",
    y = "Frequency"
  )

plot9astage
```

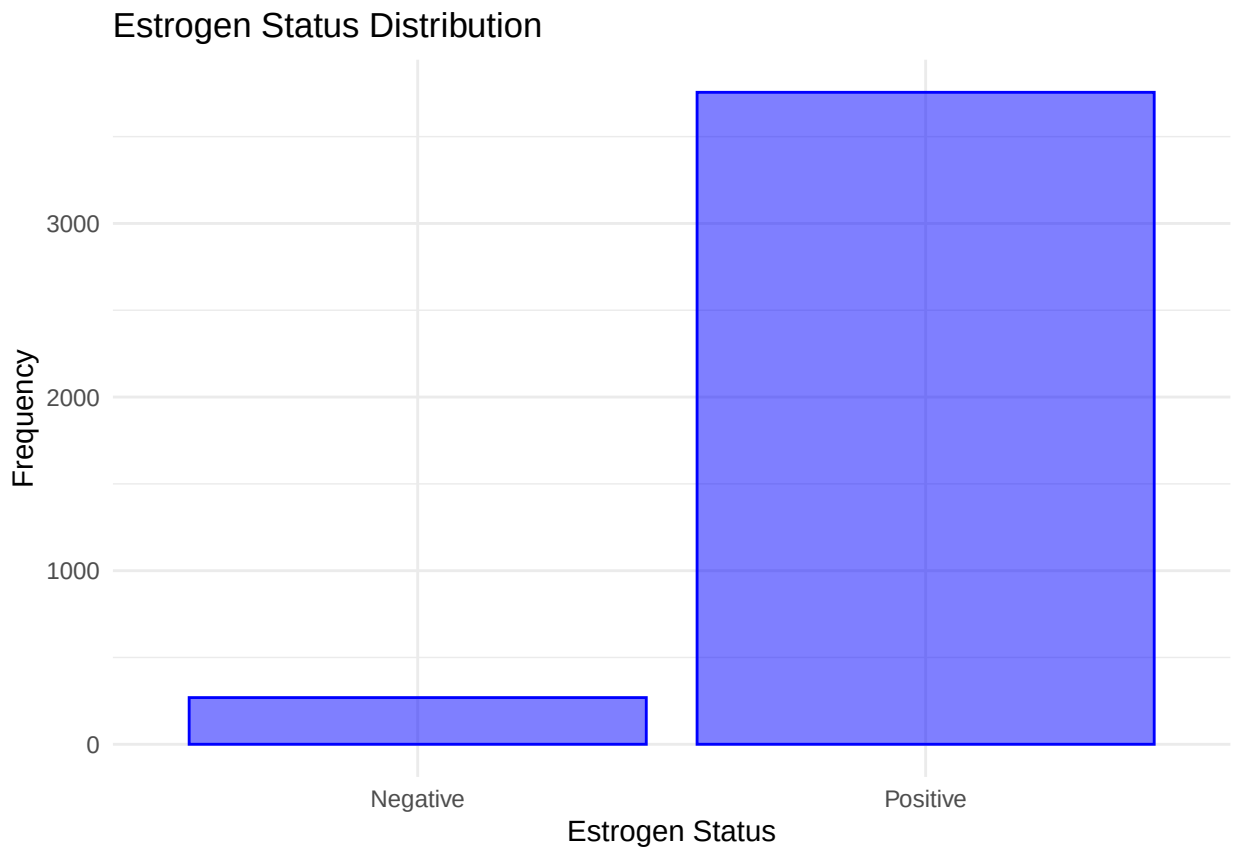



```
plot10tumorsize =
  breastcancer_data|>
  ggplot(aes(x = tumor_size)) +
  geom_histogram(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Tumor Size Distribution",
    x = "Tumor Size",
    y = "Frequency"
  )

plot10tumorsize
```

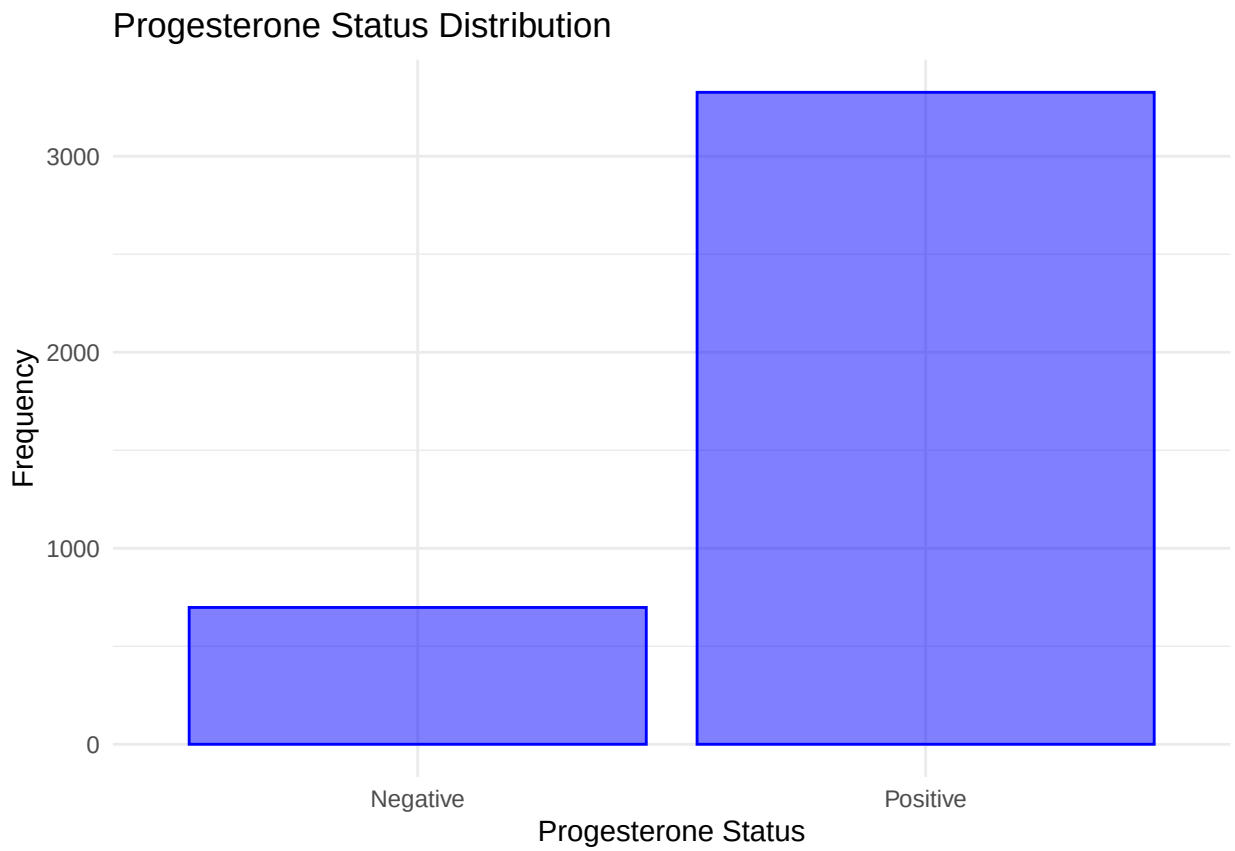
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
plot11estrogen =
  breastcancer_data|>
  ggplot(aes(x = estrogen_status)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Estrogen Status Distribution",
    x = "Estrogen Status",
    y = "Frequency"
  )

plot11estrogen
```

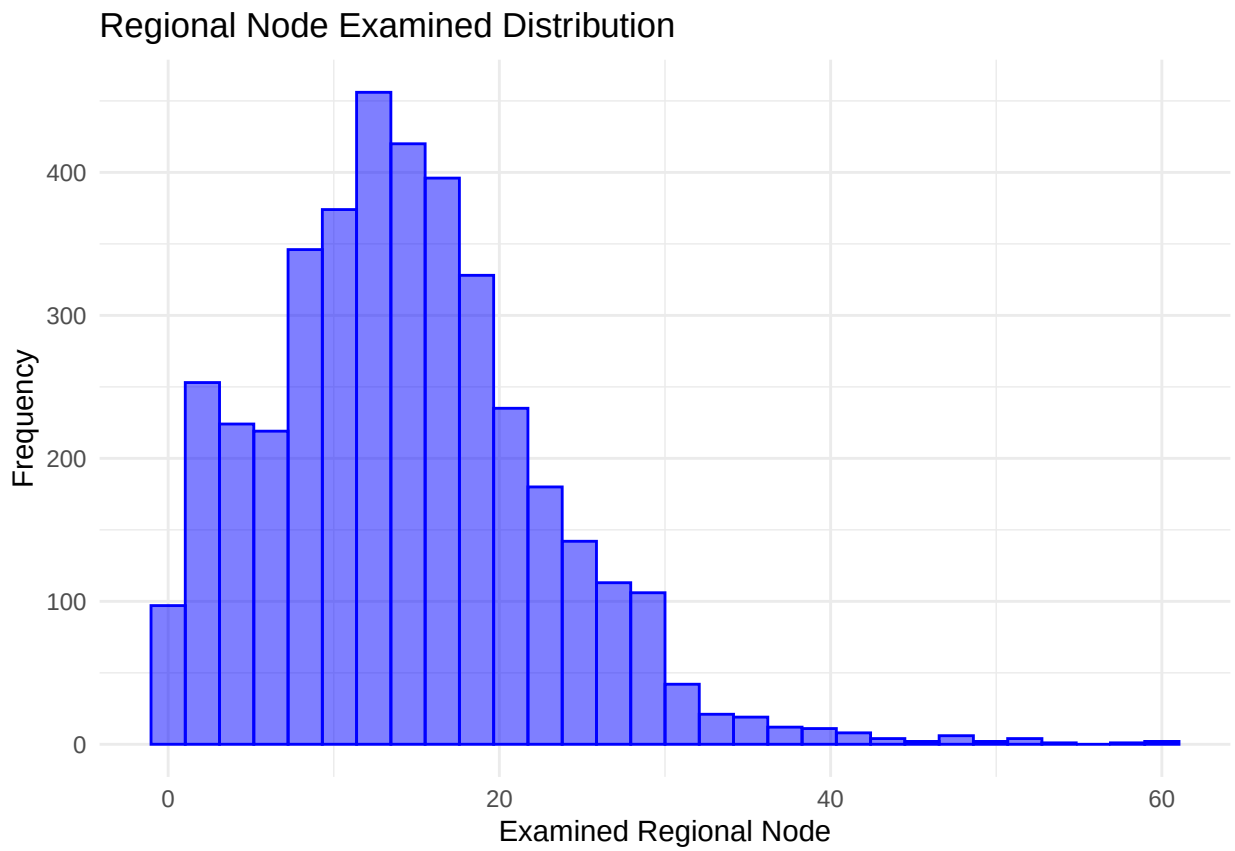



```
plot12progesterone =
  breastcancer_data|>
  ggplot(aes(x = progesterone_status)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Progesterone Status Distribution",
    x = "Progesterone Status",
    y = "Frequency"
  )

plot12progesterone
```



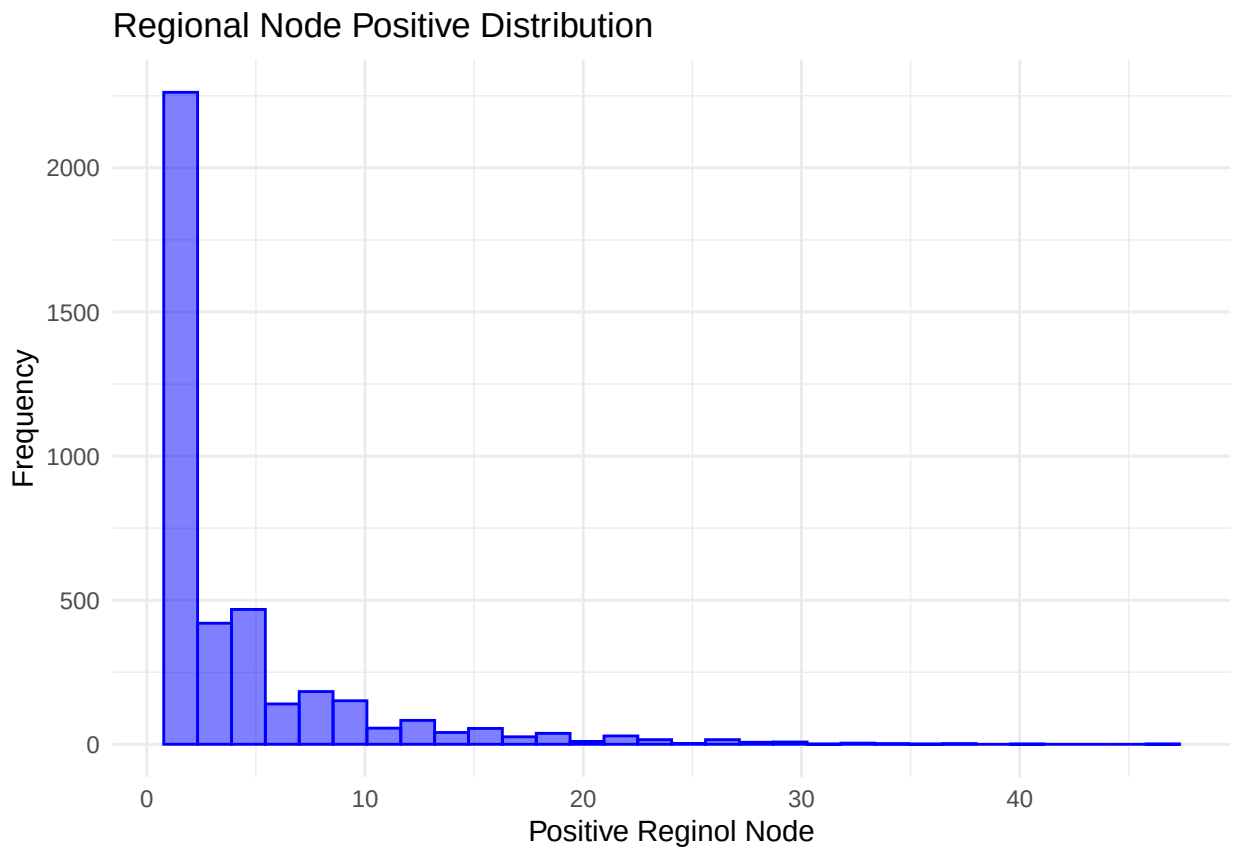

```
plot13nodeexamined =
  breastcancer_data|>
  ggplot(aes(x = regional_node_examined)) +
  geom_histogram(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Regional Node Examined Distribution",
    x = "Examined Regional Node",
    y = "Frequency"
  )

plot13nodeexamined
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
plot14nodepositive =
  breastcancer_data|>
  ggplot(aes(x = reginol_node_positive)) +
  geom_histogram(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Regional Node Positive Distribution",
    x = "Positive Reginol Node",
    y = "Frequency"
  )

plot14nodepositive
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

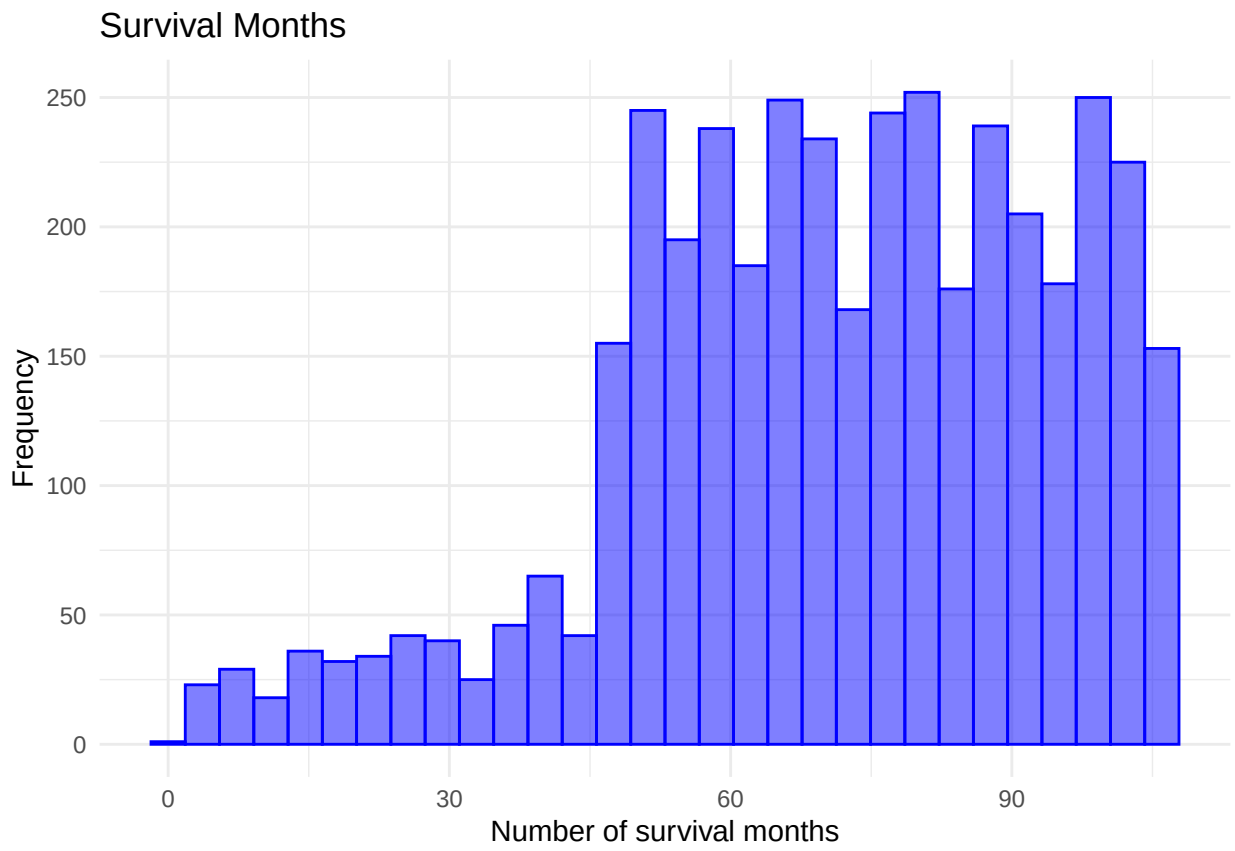



Y1 = survive months; (numeric)

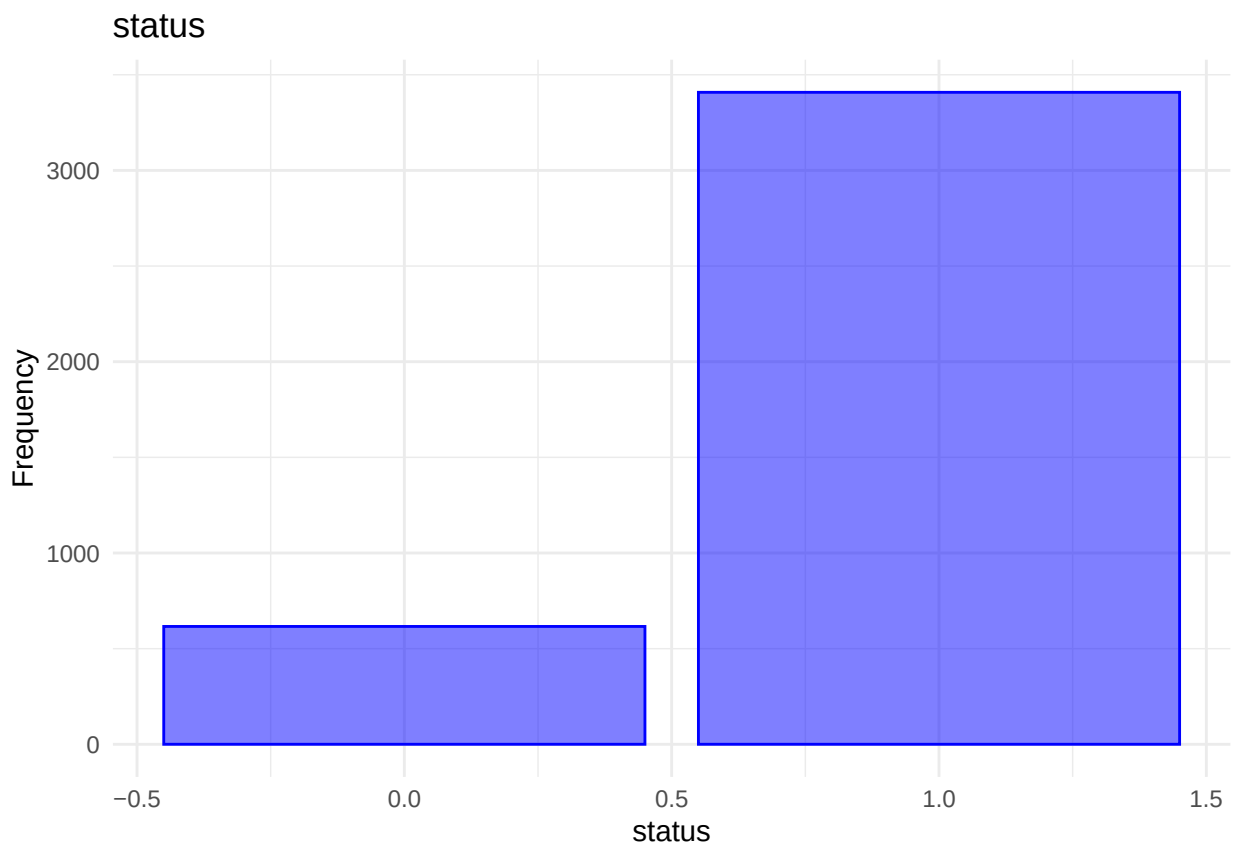
Y2 = status; (binary)

```
plot15survivalmonths =
  breastcancer_data|>
  ggplot(aes(x = survival_months)) +
  geom_histogram(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "Survival Months",
    x = "Number of survival months",
    y = "Frequency"
  )

plot15survivalmonths
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



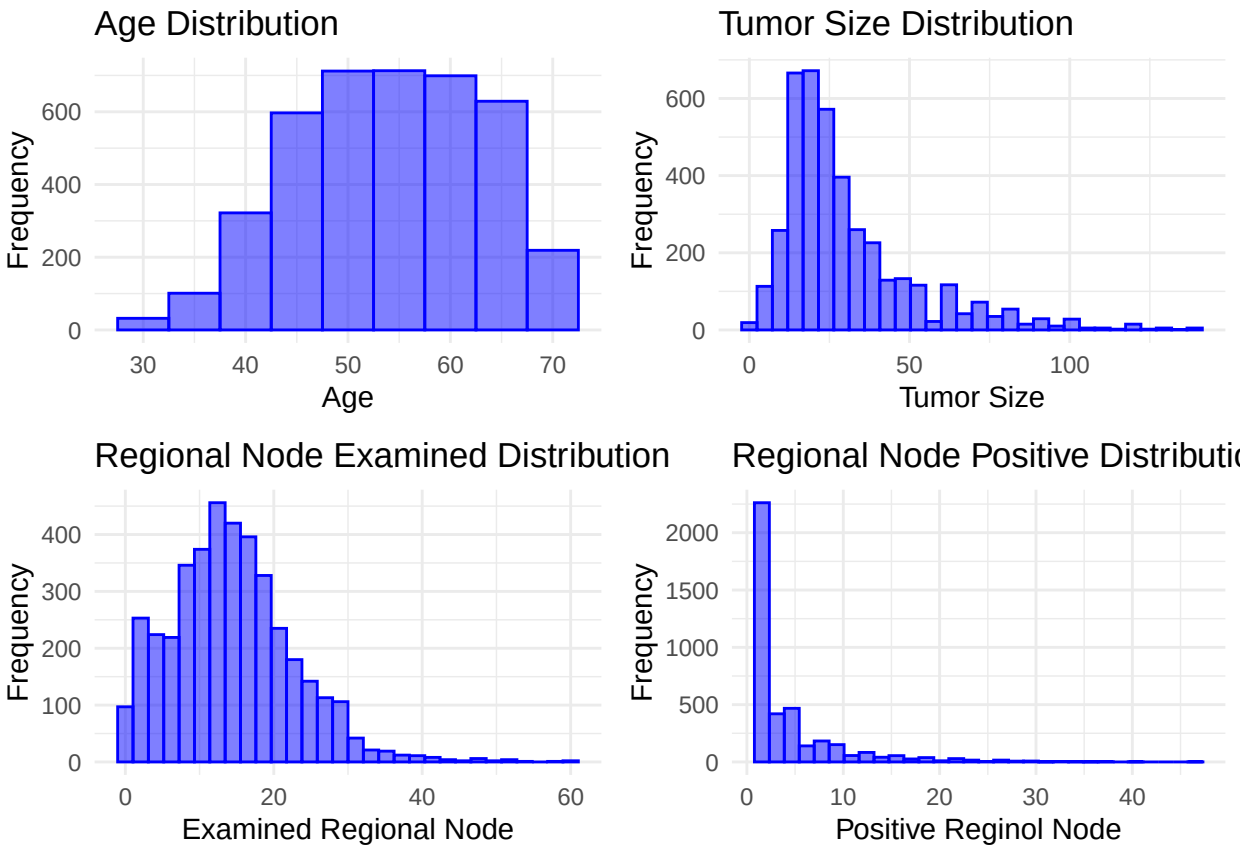

```
plot16status =
  bc|>
  ggplot(aes(x = status)) +
  geom_bar(color = "blue", fill = alpha("blue", 0.5))+
  theme_minimal() +
  labs(
    title = "status",
    x = "status",
    y = "Frequency"
  )

plot16status
```




**Summarized plots**

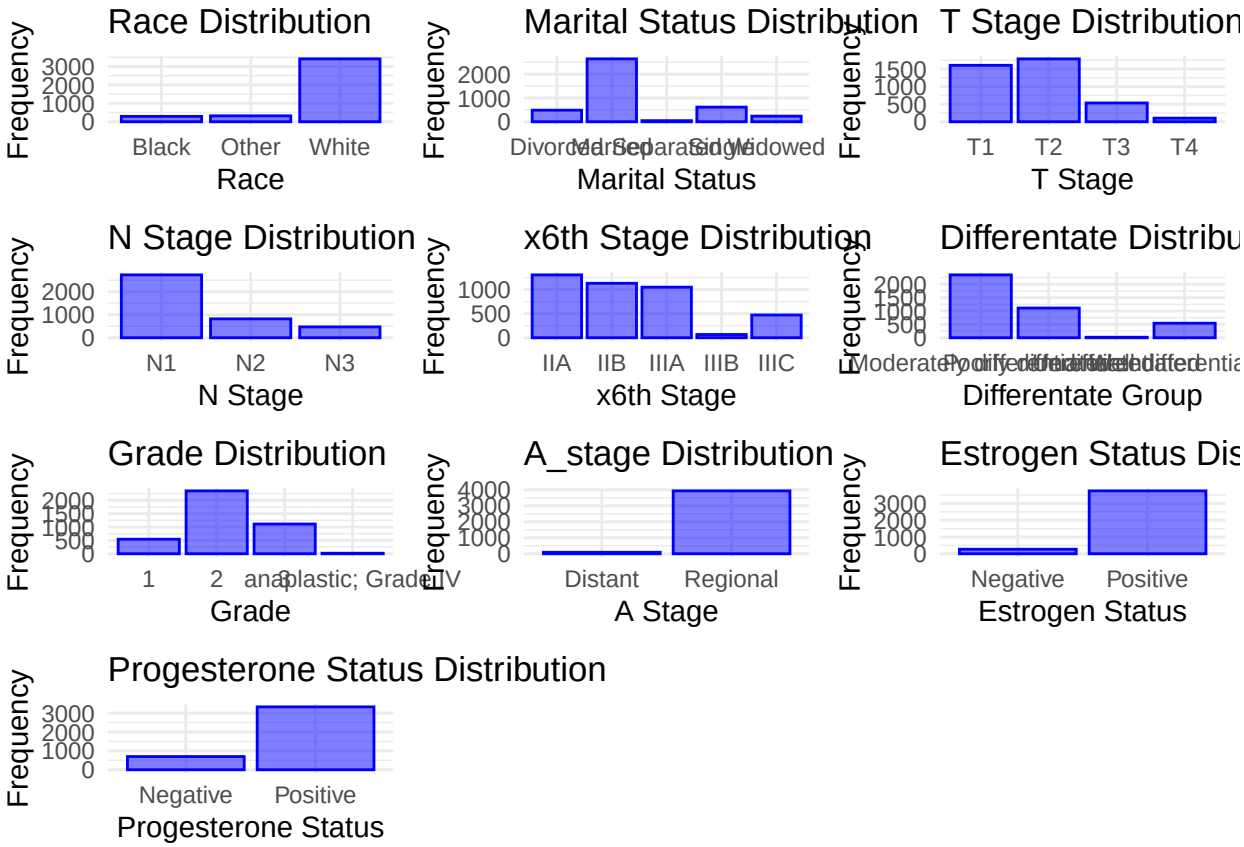
```
grid.arrange(plot1age, plot10tumorsize, plot13nodeexamined,
             plot14nodepositive, ncol = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




We can see that age is approximately normal, while tumor size, regional node examined, and regional node positive are skewed.

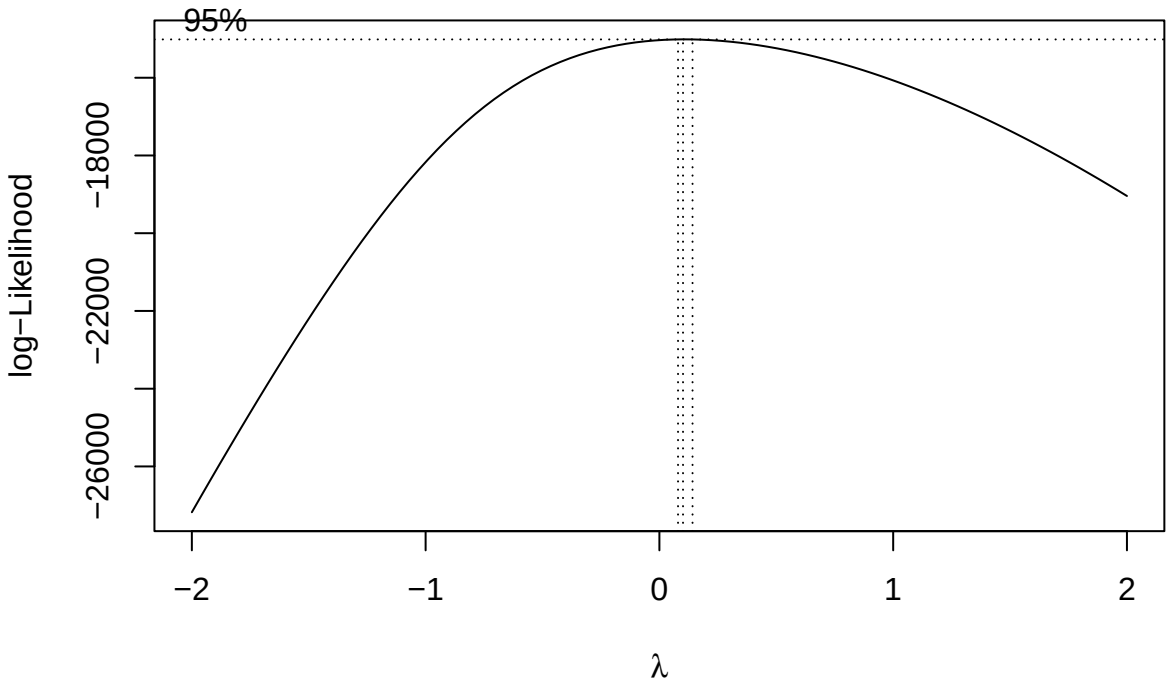
```
grid.arrange(plot2race, plot3marital,plot4tstage,
             plot5nstage, plot6x6thstage,plot7differentate,
             plot8grade,plot9astage,plot11estrogen,plot12progesterone, ncol = 3)
```



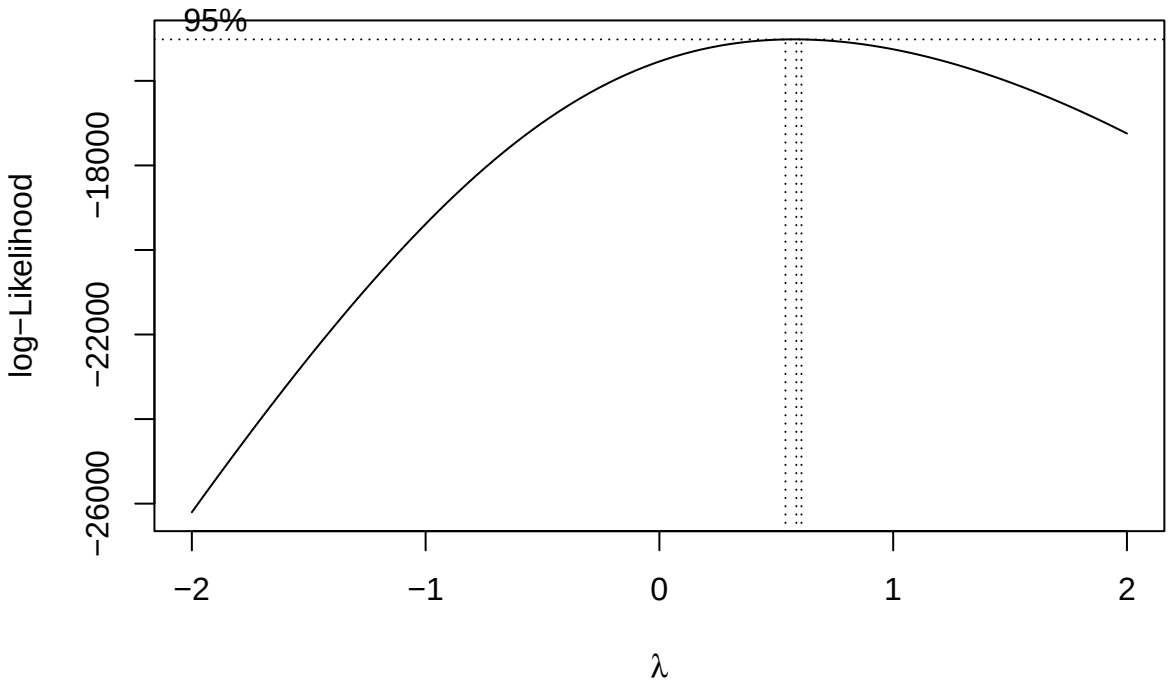

## Transformation

We know that the tumor size, regional node examined, and regional node positive are skewed. We should do transformation on these variables. Before the transformation, we can use the Box-Cox plot to check which transformation work the best for them.

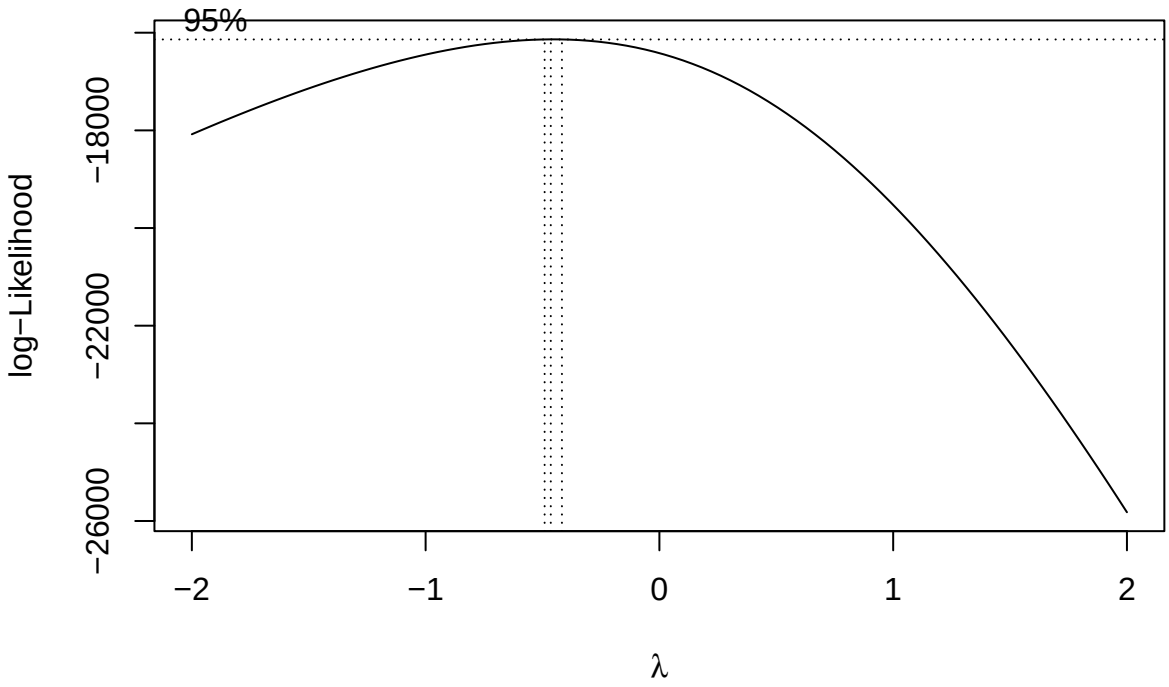
```
bc_transform_tumorsize <- boxcox(breastcancer_data$tumor_size ~ 1, lambda = seq(-2, 2, by=0.1))
```




```
bc_transform_regionalnode_examined <- boxcox(breastcancer_data$regional_node_examined ~ 1, lambda = seq
```

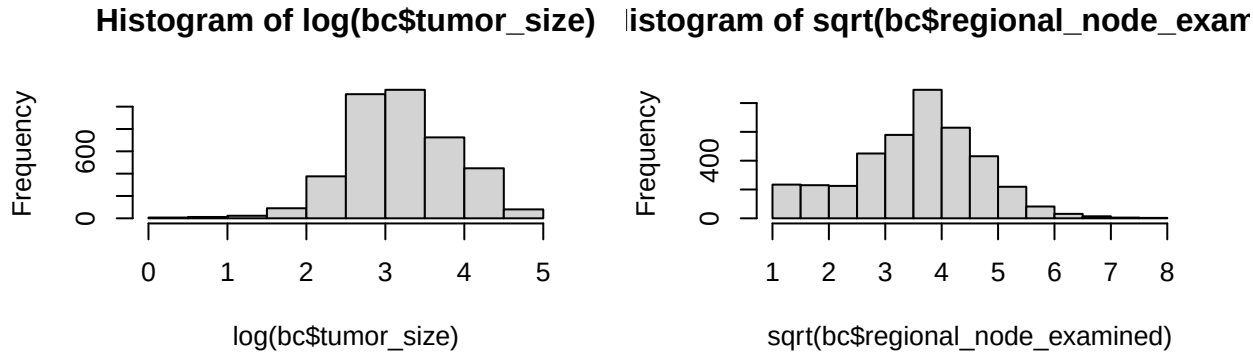


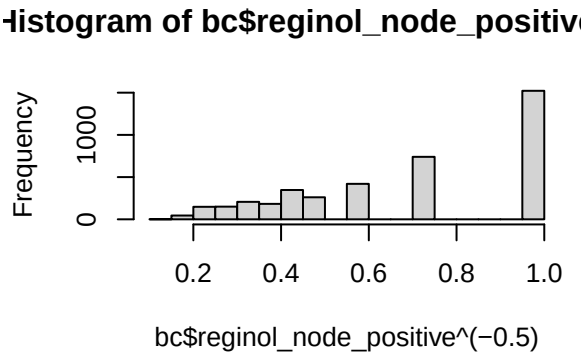

```
bc_transform_regionalnode_pos <- boxcox(breastcancer_data$reginol_node_positive ~ 1, lambda = seq(-2, 2
```




The lambda value of tumor size is close to 0, so we should use log transformation, while the lambda value of regional node examined is around 0.5, we should take a square root to the value, and the lambda value of regional node positive is around -0.5, so we should take a take square root and take an (-1) exponent for transformation.

```
par(mfrow = c(2, 2))
hist(log(bc$tumor_size))
hist(sqrt(bc$regional_node_examined))
hist(bc$reginol_node_positive**(-0.5))
```




We can see that tumor size and regional node examined become approximately normal after log transformation, while the regional node positive is still extremely skewed. Therefore, we may consider not using the variable of reginol_node_positive.

### Transformation model

```
newbc1 = bc |>
  mutate(ln_tumor=log(tumor_size),
         sqrt_examined=sqrt(regional_node_examined)) |>
  dplyr::select(-tumor_size) |>
  dplyr::select(-regional_node_examined) |>
  dplyr::select(-status)
newbc1
```

```
## # A tibble: 4,024 x 15
##      age  race marital_status t_stage n_stage x6th_stage differentiate grade
##    <dbl> <dbl>          <dbl>   <dbl>   <dbl>      <dbl>          <dbl> <dbl>
##  1    68     1              1       1       1          1              1     3
##  2    50     1              1       2       2          2              2     2
##  3    58     1              2       3       3          3              2     2
##  4    58     1              1       1       1          1              1     3
##  5    47     1              1       2       1          4              1     3
##  6    51     1              3       1       1          1              2     2
##  7    51     1              1       1       1          1              3     1
```

```
##  8    40     1              1       2       1          4             2     2
##  9    40     1              2       4       3          3             1     3
## 10    69     1              1       4       3          3             3     1
## # i 4,014 more rows
## # i 7 more variables: a_stage <dbl>, estrogen_status <dbl>,
## #   progesterone_status <dbl>, reginol_node_positive <dbl>,
## #   survival_months <dbl>, ln_tumor <dbl>, sqrt_examined <dbl>
```

```
newbc2 = bc |>
  mutate(ln_tumor=log(tumor_size),
         sqrt_examined=sqrt(regional_node_examined)) |>
  dplyr::select(-tumor_size) |>
  dplyr::select(-regional_node_examined) |>
  dplyr::select(-survival_months)
newbc2
```

```
## # A tibble: 4,024 x 15
##      age  race marital_status t_stage n_stage x6th_stage differentiate grade
##    <dbl> <dbl>          <dbl>   <dbl>   <dbl>      <dbl>         <dbl> <dbl>
##  1    68     1              1       1       1          1             1     3
##  2    50     1              1       2       2          2             2     2
##  3    58     1              2       3       3          3             2     2
##  4    58     1              1       1       1          1             1     3
##  5    47     1              1       2       1          4             1     3
##  6    51     1              3       1       1          1             2     2
##  7    51     1              1       1       1          1             3     1
##  8    40     1              1       2       1          4             2     2
##  9    40     1              2       4       3          3             1     3
## 10    69     1              1       4       3          3             3     1
## # i 4,014 more rows
## # i 7 more variables: a_stage <dbl>, estrogen_status <dbl>,
## #   progesterone_status <dbl>, reginol_node_positive <dbl>, status <dbl>,
## #   ln_tumor <dbl>, sqrt_examined <dbl>
```

## Indicator Test

### When y is status

```
# indicator test when y is status
categorical_vars <- c("race", "marital_status", "t_stage", "n_stage", "x6th_stage",
                      "differentiate", "grade", "a_stage",
                      "estrogen_status", "progesterone_status")

newbc2[categorical_vars] <- lapply(newbc2[categorical_vars], factor)

formula <- as.formula("status ~ race + marital_status + t_stage + n_stage + x6th_stage +
                      differentiate + grade + a_stage + estrogen_status + progesterone_status+ln_tumor

model <- glm(formula, data = newbc2, family = binomial())

summary(model)
```

```
##
## Call:
## glm(formula = formula, family = binomial(), data = newbc2)
##
## Coefficients: (4 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.826084   0.591231   3.089  0.00201 **
## race2                  -0.517829   0.161866  -3.199  0.00138 **
## race3                   0.412988   0.202323   2.041  0.04123 *
## marital_status2        -0.208737   0.141798  -1.472  0.14100
## marital_status3        -0.137065   0.134860  -1.016  0.30946
## marital_status4        -0.226852   0.192395  -1.179  0.23836
## marital_status5        -0.870139   0.369406  -2.356  0.01850 *
## t_stage2               -0.241275   0.214882  -1.123  0.26151
## t_stage3               -0.463184   0.308144  -1.503  0.13280
## t_stage4               -0.911657   0.451201  -2.021  0.04333 *
## n_stage2               -0.652606   0.238058  -2.741  0.00612 **
## n_stage3               -0.757283   0.301072  -2.515  0.01189 *
## x6th_stage2             0.069464   0.294083   0.236  0.81327
## x6th_stage3                   NA         NA      NA       NA
## x6th_stage4            -0.226947   0.231875  -0.979  0.32771
## x6th_stage5            -0.085418   0.528445  -0.162  0.87159
## differentiate2          0.387813   0.104972   3.694  0.00022 ***
## differentiate3          0.922456   0.193026   4.779 1.76e-06 ***
## differentiate4         -0.970582   0.533726  -1.819  0.06899 .
## grade2                        NA         NA      NA       NA
## grade3                        NA         NA      NA       NA
## grade4                        NA         NA      NA       NA
## a_stage1                0.044840   0.266049   0.169  0.86616
## estrogen_status1        0.733142   0.177768   4.124 3.72e-05 ***
## progesterone_status1    0.589919   0.127619   4.623 3.79e-06 ***
## ln_tumor               -0.053874   0.138931  -0.388  0.69818
## sqrt_examined           0.256137   0.049748   5.149 2.62e-07 ***
## reginol_node_positive -0.074309   0.015040  -4.941 7.78e-07 ***
## age                    -0.023985   0.005625  -4.264 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2951.6  on 3999  degrees of freedom
## AIC: 3001.6
##
## Number of Fisher Scoring iterations: 5
```

Based on the above indicator test summary, we delete grade and x6th_stage because their output was NA in the output linear model, since NA indicates these predicts may contribute collinearity.

**When y is Survival Months**

```r
# indicator test when y is Survival Months
categorical_vars <- c("race", "marital_status", "t_stage", "n_stage", "x6th_stage",
                      "differentiate", "grade", "a_stage",
                      "estrogen_status", "progesterone_status")

newbc1[categorical_vars] <- lapply(newbc1[categorical_vars], factor)
formula1 <- as.formula("survival_months ~ race + marital_status + t_stage + n_stage + x6th_stage +
                      differentiate + grade + a_stage + estrogen_status + progesterone_status+ln_tumor

model1 <- lm(formula1, data = newbc1)

summary(model1)
```

```
##
## Call:
## lm(formula = formula1, data = newbc1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.557 -15.564   1.209  18.142  56.447
##
## Coefficients: (4 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              62.92531    4.78147  13.160  < 2e-16 ***
## race2                    -3.75385    1.40240  -2.677  0.00746 **
## race3                     1.90888    1.32316   1.443  0.14919
## marital_status2          -0.71596    1.11597  -0.642  0.52119
## marital_status3          -0.70625    1.02666  -0.688  0.49155
## marital_status4          -1.48374    1.57047  -0.945  0.34483
## marital_status5          -6.94522    3.40072  -2.042  0.04119 *
## t_stage2                 -1.71846    1.78646  -0.962  0.33614
## t_stage3                 -0.79270    2.51886  -0.315  0.75300
## t_stage4                 -3.38505    4.42219  -0.765  0.44404
## n_stage2                 -0.92483    1.97717  -0.468  0.63999
## n_stage3                 -3.46678    2.67103  -1.298  0.19439
## x6th_stage2              -0.41077    2.35318  -0.175  0.86143
## x6th_stage3                    NA         NA      NA       NA
## x6th_stage4               0.57306    1.82486   0.314  0.75351
## x6th_stage5               3.64333    5.15554   0.707  0.47980
## differentiate2            0.99136    0.85199   1.164  0.24466
## differentiate3            0.93301    1.22642   0.761  0.44685
## differentiate4           -1.88819    5.23011  -0.361  0.71810
## grade2                         NA         NA      NA       NA
## grade3                         NA         NA      NA       NA
## grade4                         NA         NA      NA       NA
## a_stage1                  4.36402    2.67234   1.633  0.10254
## estrogen_status1          8.62044    1.68608   5.113 3.32e-07 ***
## progesterone_status1      1.64605    1.10093   1.495  0.13496
## ln_tumor                 -1.10747    1.02996  -1.075  0.28233
## sqrt_examined             0.81816    0.34568   2.367  0.01799 *
## reginol_node_positive    -0.31002    0.14135  -2.193  0.02835 *
## age                      -0.04076    0.04140  -0.985  0.32493
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.47 on 3999 degrees of freedom
## Multiple R-squared:  0.04452,    Adjusted R-squared:  0.03879
## F-statistic: 7.764 on 24 and 3999 DF,  p-value: < 2.2e-16
```

Based on the above indicator test summary, we delete grade and x6th_stage.

## Model Fitting

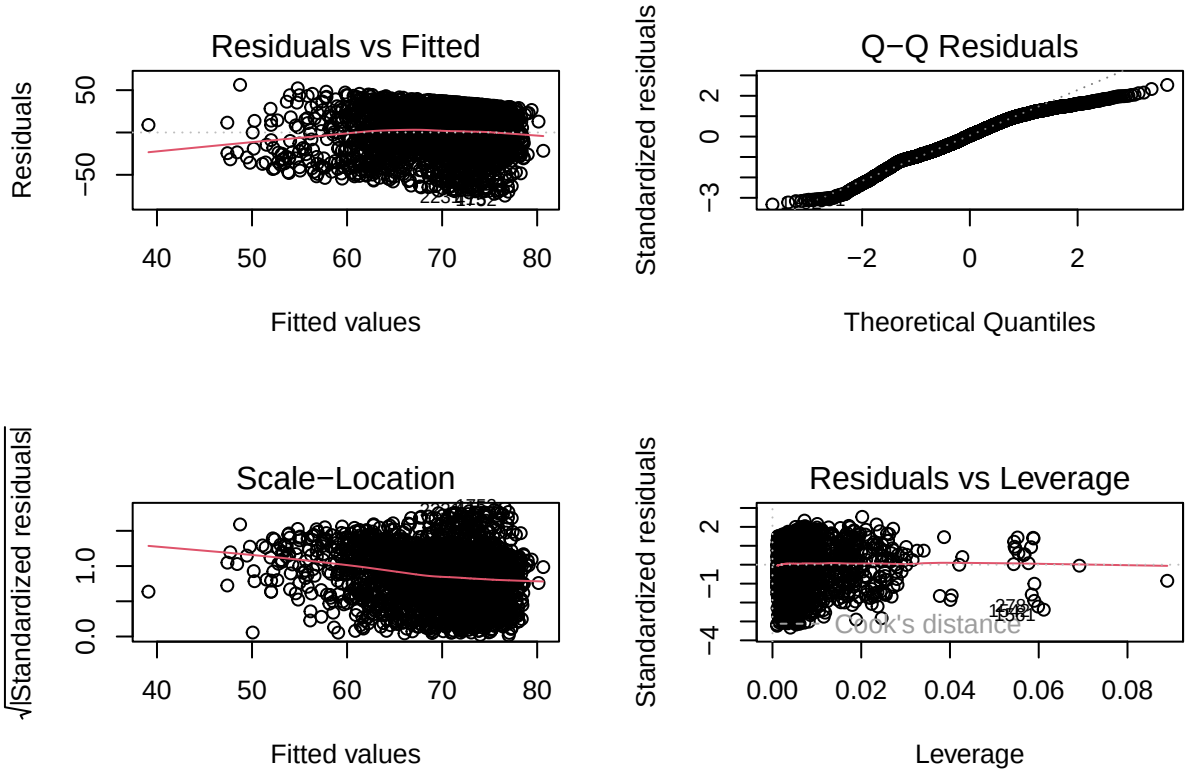
### Initial Model

```
lmfit=lm(survival_months ~ race + marital_status + t_stage + n_stage + differentiate + a_stage + estrog
summary(lmfit)
```

```
##
## Call:
## lm(formula = survival_months ~ race + marital_status + t_stage +
##     n_stage + differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + reginol_node_positive + age, data = newbc1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.609 -15.585   1.257  18.080  56.256
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              63.17078    4.76959  13.245  < 2e-16 ***
## race2                    -3.72212    1.40154  -2.656  0.00794 **
## race3                     1.90515    1.32220   1.441  0.14969
## marital_status2          -0.71759    1.11560  -0.643  0.52011
## marital_status3          -0.71874    1.02534  -0.701  0.48336
## marital_status4          -1.49914    1.56989  -0.955  0.33967
## marital_status5          -7.06187    3.39685  -2.079  0.03769 *
## t_stage2                 -1.29191    1.14280  -1.130  0.25834
## t_stage3                 -0.86768    2.01176  -0.431  0.66627
## t_stage4                 -0.71612    2.78843  -0.257  0.79733
## n_stage2                 -1.54282    1.06924  -1.443  0.14912
## n_stage3                 -3.94737    2.23622  -1.765  0.07761 .
## differentiate2            1.00442    0.85141   1.180  0.23818
## differentiate3            0.95437    1.22513   0.779  0.43603
## differentiate4           -2.15661    5.21968  -0.413  0.67950
## a_stage1                  4.27338    2.66881   1.601  0.10940
## estrogen_status1          8.57571    1.68488   5.090 3.75e-07 ***
## progesterone_status1      1.65003    1.10050   1.499  0.13386
## ln_tumor                 -1.12285    1.02902  -1.091  0.27526
## sqrt_examined             0.82002    0.34542   2.374  0.01765 *
## reginol_node_positive    -0.30629    0.14084  -2.175  0.02971 *
## age                      -0.04166    0.04135  -1.007  0.31377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 22.47 on 4002 degrees of freedom
## Multiple R-squared:  0.04432,    Adjusted R-squared:  0.03931
## F-statistic: 8.838 on 21 and 4002 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lmfit)
```

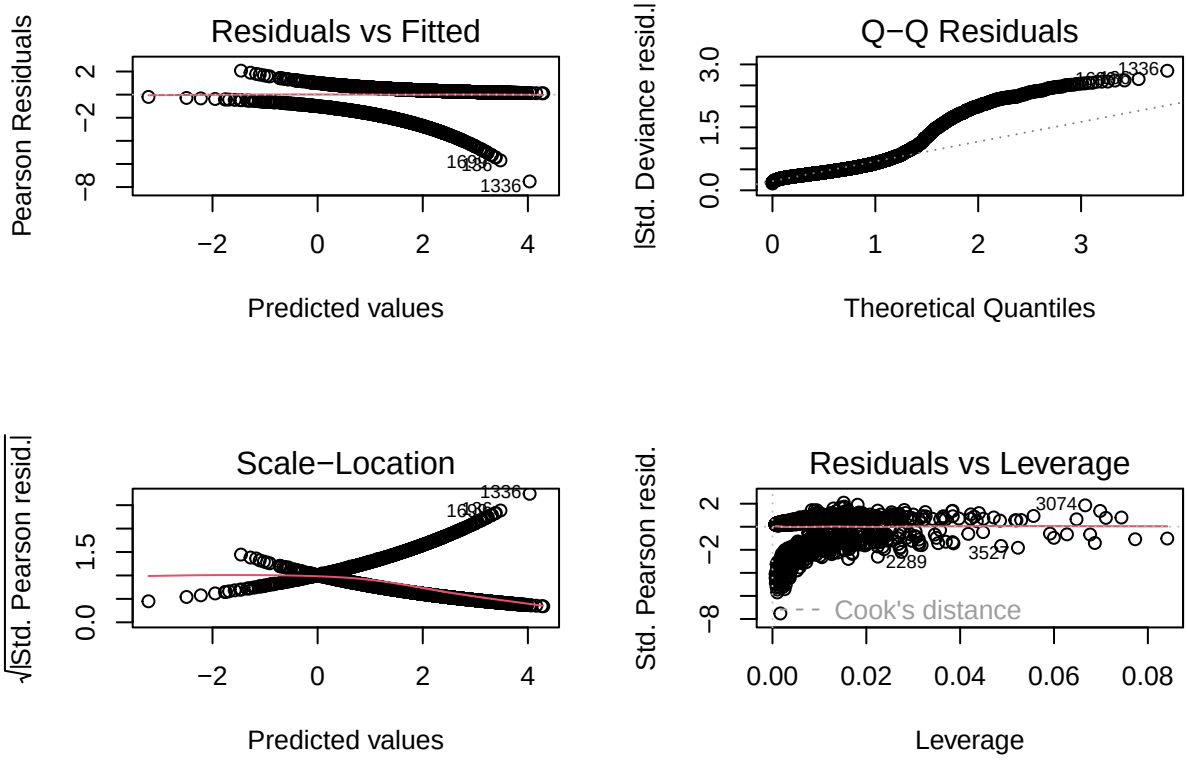



```
glmfit <- glm(status ~ race + marital_status + t_stage + n_stage + differentiate +
                a_stage + estrogen_status + progesterone_status + ln_tumor +
                sqrt_examined + reginol_node_positive + age,
                data = newbc2, family = binomial)
summary(glmfit)
```

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + reginol_node_positive + age, family = binomial,
##     data = newbc2)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              1.775787   0.587386   3.023 0.002501 **
## race2                   -0.522765   0.161744  -3.232 0.001229 **
```

```
## race3                  0.419131   0.202312   2.072 0.038293 *
## marital_status2       -0.209325   0.141735  -1.477 0.139709
## marital_status3       -0.140584   0.134542  -1.045 0.296064
## marital_status4       -0.222113   0.192496  -1.154 0.248558
## marital_status5       -0.867369   0.370116  -2.344 0.019104 *
## t_stage2              -0.372488   0.159129  -2.341 0.019243 *
## t_stage3              -0.471905   0.266764  -1.769 0.076894 .
## t_stage4              -1.025962   0.313035  -3.277 0.001047 **
## n_stage2              -0.474091   0.129288  -3.667 0.000245 ***
## n_stage3              -0.637101   0.237639  -2.681 0.007341 **
## differentiate2         0.386027   0.104898   3.680 0.000233 ***
## differentiate3         0.917457   0.192668   4.762 1.92e-06 ***
## differentiate4        -0.947095   0.531127  -1.783 0.074557 .
## a_stage1               0.045211   0.265718   0.170 0.864894
## estrogen_status1       0.737828   0.177538   4.156 3.24e-05 ***
## progesterone_status1   0.588385   0.127499   4.615 3.93e-06 ***
## ln_tumor              -0.055274   0.138631  -0.399 0.690103
## sqrt_examined          0.255997   0.049654   5.156 2.53e-07 ***
## reginol_node_positive -0.074961   0.014984  -5.003 5.65e-07 ***
## age                   -0.023647   0.005618  -4.209 2.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2953.4  on 4002  degrees of freedom
## AIC: 2997.4
##
## Number of Fisher Scoring iterations: 5
```

```
par(mfrow = c(2, 2))
plot(glmfit)
```




## Partial Test

### Partial test for numeric Y

```
# Our 1st global model for the predicts without collinearity and not normal predicts

model_global_num = lm(survival_months ~ race + marital_status + t_stage + n_stage +
                        differentiate + a_stage + estrogen_status + progesterone_status+ln_tumor + sqrt_

summary(model_global_num)
```

```
##
## Call:
## lm(formula = survival_months ~ race + marital_status + t_stage +
##     n_stage + differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + age, data = newbc1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.099 -15.622   1.257  18.199  53.466
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)          63.72848    4.76490  13.375  < 2e-16 ***
## race2                -3.64322    1.40172  -2.599  0.00938 **
## race3                 1.89824    1.32282   1.435  0.15137
## marital_status2      -0.72900    1.11611  -0.653  0.51369
## marital_status3      -0.73946    1.02577  -0.721  0.47102
## marital_status4      -1.48399    1.57061  -0.945  0.34479
## marital_status5      -7.37323    3.39542  -2.172  0.02995 *
## t_stage2             -1.29134    1.14333  -1.129  0.25877
## t_stage3             -0.85478    2.01269  -0.425  0.67108
## t_stage4             -0.87978    2.78871  -0.315  0.75241
## n_stage2             -2.68831    0.93094  -2.888  0.00390 **
## n_stage3             -7.94856    1.27167  -6.250 4.52e-10 ***
## differentiate2        0.98469    0.85176   1.156  0.24772
## differentiate3        0.92668    1.22564   0.756  0.44964
## differentiate4       -2.15317    5.22211  -0.412  0.68013
## a_stage1              4.06226    2.66828   1.522  0.12798
## estrogen_status1      8.57788    1.68567   5.089 3.77e-07 ***
## progesterone_status1  1.63586    1.10099   1.486  0.13741
## ln_tumor             -1.16315    1.02933  -1.130  0.25854
## sqrt_examined         0.64625    0.33621   1.922  0.05466 .
## age                  -0.04416    0.04135  -1.068  0.28563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.48 on 4003 degrees of freedom
## Multiple R-squared:  0.04319,    Adjusted R-squared:  0.03841
## F-statistic: 9.035 on 20 and 4003 DF,  p-value: < 2.2e-16
```

From the summary of the global model, we can see that from the P-value of race, t_stage, differentiate, a_stage, progesterone_status, ln_tumor, sqrt_examined, and age all have the p-value above 0.05 which is not significant. While sqrt_examined have a 0.0559 p-value which is a close call, so we may consider keep it. Since the covariates like marital_status, n_stage, estrogen_status have a P-value smaller than 0.05 indicating significance. Therefore, our 1st partial test may include the significant covariates.

```
# 1st numeric partial test with all significant covariates
nummodel_partial_1 = lm(survival_months ~ marital_status + n_stage + estrogen_status + sqrt_examined, ne

summary(nummodel_partial_1)
```

```
##
## Call:
## lm(formula = survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined, data = newbc1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.161 -15.683   1.245  18.270  52.520
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       61.6132     1.8481  33.339  < 2e-16 ***
## marital_status2   -1.1696     1.1130  -1.051 0.293396
## marital_status3   -1.3670     1.0091  -1.355 0.175583
```

```
## marital_status4   -2.1762     1.5344  -1.418 0.156189
## marital_status5   -8.0320     3.3909  -2.369 0.017898 *
## n_stage2          -3.3926     0.9133  -3.715 0.000206 ***
## n_stage3          -9.6707     1.1909  -8.120 6.13e-16 ***
## estrogen_status1  10.4829     1.4306   7.327 2.82e-13 ***
## sqrt_examined      0.6343     0.3355   1.891 0.058733 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.52 on 4015 degrees of freedom
## Multiple R-squared:  0.03623,    Adjusted R-squared:  0.03431
## F-statistic: 18.87 on 8 and 4015 DF,  p-value: < 2.2e-16
```

```
# 2nd numeric partial test with all in-significant covariates
nummodel_partial_2 = lm(survival_months ~ race + t_stage + differentiate + a_stage + progesterone_statu

summary(nummodel_partial_2)
```

```
##
## Call:
## lm(formula = survival_months ~ race + t_stage + differentiate +
##     a_stage + progesterone_status + ln_tumor + age, data = newbc1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.014 -15.237   1.229  18.385  49.728
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          66.19189    4.50616  14.689  < 2e-16 ***
## race2                -4.43559    1.39084  -3.189 0.001438 **
## race3                 1.80738    1.33187   1.357 0.174848
## t_stage2             -1.51866    1.15089  -1.320 0.187061
## t_stage3             -1.33650    2.02590  -0.660 0.509480
## t_stage4             -1.54160    2.80670  -0.549 0.582860
## differentiate2        2.06327    0.84699   2.436 0.014894 *
## differentiate3        2.17977    1.22124   1.785 0.074355 .
## differentiate4       -4.14026    5.25692  -0.788 0.430988
## a_stage1              8.85815    2.58816   3.423 0.000627 ***
## progesterone_status1  4.86565    0.96354   5.050 4.62e-07 ***
## ln_tumor             -1.66881    1.03323  -1.615 0.106358
## age                  -0.04877    0.04041  -1.207 0.227487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.67 on 4011 degrees of freedom
## Multiple R-squared:  0.02513,    Adjusted R-squared:  0.02221
## F-statistic: 8.617 on 12 and 4011 DF,  p-value: < 2.2e-16
```

```
# 3rd numeric partial test with some significant in 2nd test added
nummodel_partial_3 = lm(survival_months ~ marital_status + n_stage + estrogen_status + sqrt_examined + a

summary(nummodel_partial_3)
```

```
##
## Call:
## lm(formula = survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined + a_stage + progesterone_status + ln_tumor,
##     data = newbc1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.751 -15.561   1.111  18.340  53.113
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            62.7481     3.5394  17.729  < 2e-16 ***
## marital_status2        -1.0285     1.1121  -0.925  0.35509
## marital_status3        -1.2491     1.0082  -1.239  0.21546
## marital_status4        -2.1415     1.5331  -1.397  0.16254
## marital_status5        -7.8368     3.3883  -2.313  0.02078 *
## n_stage2               -2.8026     0.9282  -3.019  0.00255 **
## n_stage3               -8.2702     1.2634  -6.546 6.66e-11 ***
## estrogen_status1        8.8999     1.6608   5.359 8.85e-08 ***
## sqrt_examined           0.6575     0.3355   1.960  0.05010 .
## a_stage1                3.9132     2.4926   1.570  0.11650
## progesterone_status1    1.8102     1.0946   1.654  0.09826 .
## ln_tumor               -1.6793     0.5603  -2.997  0.00274 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.49 on 4012 degrees of freedom
## Multiple R-squared:  0.03978,    Adjusted R-squared:  0.03715
## F-statistic: 15.11 on 11 and 4012 DF,  p-value: < 2.2e-16
```

- Our 1st partial test only include the significant predicts on global test, and then we also did a second partial test includes all the covariates that are non significant, and we can see the 1st partial model has an adjusted R-squared of 0.03419 which is slightly lower than the global test with the same RSE to global test indicating a slightly underfitting. While the second model with all the non signifcant covraites only get an adjusted R-squared of 0.01859 which is much lower than the global test with a higher RSE, so we will definitely not use the model 2 combination. However, we do discover some covariates are significant on the second model like n_stage, and progesterone_status, so we include these 3 again onto the 1st model to get our 3rd model. Fortunately, our 3rd model has a highest adjusted R-squared value of 0.03714 and a lower RSE.

### Partial Test for binary Y

```
# Our 1st global model for the predicts without colinearity and not normal predicts
model_global_bin = glm(status ~ race + marital_status + t_stage + n_stage +
                       differentiate + a_stage +estrogen_status + progesterone_status + ln_tumor + sqrt

summary(model_global_bin)
```

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
```

```
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + age, family = binomial, data = newbc2)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.025431   0.583766   3.470 0.000521 ***
## race2                -0.490040   0.161448  -3.035 0.002403 **
## race3                 0.412102   0.201264   2.048 0.040602 *
## marital_status2      -0.213770   0.140919  -1.517 0.129273
## marital_status3      -0.149790   0.134179  -1.116 0.264275
## marital_status4      -0.220417   0.191076  -1.154 0.248681
## marital_status5      -0.936900   0.365551  -2.563 0.010378 *
## t_stage2             -0.355769   0.158618  -2.243 0.024902 *
## t_stage3             -0.450653   0.266574  -1.691 0.090925 .
## t_stage4             -1.044577   0.312476  -3.343 0.000829 ***
## n_stage2             -0.733552   0.117910  -6.221 4.93e-10 ***
## n_stage3             -1.575446   0.141592 -11.127  < 2e-16 ***
## differentiate2        0.375295   0.104473   3.592 0.000328 ***
## differentiate3        0.901142   0.192252   4.687 2.77e-06 ***
## differentiate4       -0.960780   0.535816  -1.793 0.072955 .
## a_stage1              0.002227   0.262767   0.008 0.993239
## estrogen_status1      0.730261   0.176959   4.127 3.68e-05 ***
## progesterone_status1  0.573254   0.127304   4.503 6.70e-06 ***
## ln_tumor             -0.065530   0.138852  -0.472 0.636966
## sqrt_examined         0.180931   0.046777   3.868 0.000110 ***
## age                  -0.024590   0.005591  -4.398 1.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2978.6  on 4003  degrees of freedom
## AIC: 3020.6
##
## Number of Fisher Scoring iterations: 5
```

From the summary of the global model, we can see that from the P-value of a_stage, ln_tumor, marital_status(2, 3, 4) have the p-value above 0.05 which is not significant. While differentiate has a 0.0729 p-value, and t_stage3 has a p-value of 0.0909, we may consider keep them. The covariates like t_stage, n_stage, race, estrogen_status, progesterone_status, sqrt_examined, age have a P-value smaller than 0.05 indicating significance. Therefore, our 1st partial test may include the significant covariates.

```
# 1st binary partial test with all significant covariates
# We will still keep marital_status because one group in this variable is significant
binmodel_partial_1 = glm(status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_s

summary(binmodel_partial_1)
```

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + estrogen_status + progesterone_status + sqrt_examined +
```

```
##     age, family = binomial, data = newbc2)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.851463   0.377142   4.909 9.15e-07 ***
## race2                -0.488627   0.161358  -3.028 0.002460 **
## race3                 0.411208   0.201230   2.043 0.041006 *
## marital_status2      -0.215703   0.140825  -1.532 0.125594
## marital_status3      -0.151385   0.134089  -1.129 0.258903
## marital_status4      -0.221820   0.191055  -1.161 0.245631
## marital_status5      -0.938507   0.365357  -2.569 0.010207 *
## t_stage2             -0.408664   0.112638  -3.628 0.000285 ***
## t_stage3             -0.555367   0.148281  -3.745 0.000180 ***
## t_stage4             -1.125881   0.244474  -4.605 4.12e-06 ***
## n_stage2             -0.737853   0.117555  -6.277 3.46e-10 ***
## n_stage3             -1.581593   0.137038 -11.541  < 2e-16 ***
## differentiate2        0.377070   0.104385   3.612 0.000303 ***
## differentiate3        0.904754   0.191998   4.712 2.45e-06 ***
## differentiate4       -0.961568   0.535510  -1.796 0.072557 .
## estrogen_status1      0.730173   0.176789   4.130 3.62e-05 ***
## progesterone_status1  0.573495   0.127268   4.506 6.60e-06 ***
## sqrt_examined         0.181071   0.046740   3.874 0.000107 ***
## age                  -0.024489   0.005585  -4.385 1.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2978.9  on 4005  degrees of freedom
## AIC: 3016.9
##
## Number of Fisher Scoring iterations: 5
```

```
# 2nd binary partial test with all in-significant covariates
binmodel_partial_2 = glm(status ~ a_stage + ln_tumor, family = binomial, newbc2)

summary(binmodel_partial_2)
```

```
##
## Call:
## glm(formula = status ~ a_stage + ln_tumor, family = binomial,
##     data = newbc2)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.53355    0.34110   7.428 1.11e-13 ***
## a_stage1     1.02938    0.22401   4.595 4.32e-06 ***
## ln_tumor    -0.55245    0.07032  -7.856 3.97e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 3351.7  on 4021  degrees of freedom
## AIC: 3357.7
##
## Number of Fisher Scoring iterations: 4
```

```
# 3rd binary partial test with some significant in 2nd test added
binmodel_partial_3 = glm(status ~ marital_status + t_stage + n_stage + differentiate + estrogen_status +

summary(binmodel_partial_3)
```

```
##
## Call:
## glm(formula = status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     a_stage + ln_tumor, family = binomial, data = newbc2)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            1.985848   0.582219   3.411 0.000648 ***
## marital_status2       -0.247509   0.140237  -1.765 0.077575 .
## marital_status3       -0.231808   0.131581  -1.762 0.078119 .
## marital_status4       -0.281274   0.189215  -1.487 0.137138
## marital_status5       -0.986920   0.363073  -2.718 0.006563 **
## t_stage2              -0.346774   0.158600  -2.186 0.028781 *
## t_stage3              -0.450100   0.267133  -1.685 0.092003 .
## t_stage4              -1.056179   0.311620  -3.389 0.000701 ***
## n_stage2              -0.736774   0.117479  -6.272 3.57e-10 ***
## n_stage3              -1.584207   0.141279 -11.213  < 2e-16 ***
## differentiate2         0.390591   0.104008   3.755 0.000173 ***
## differentiate3         0.920784   0.191675   4.804 1.56e-06 ***
## differentiate4        -0.993288   0.534044  -1.860 0.062894 .
## estrogen_status1       0.731365   0.176189   4.151 3.31e-05 ***
## progesterone_status1   0.573479   0.126984   4.516 6.30e-06 ***
## sqrt_examined          0.183193   0.046609   3.930 8.48e-05 ***
## age                   -0.024564   0.005550  -4.426 9.60e-06 ***
## a_stage1              -0.004192   0.262185  -0.016 0.987243
## ln_tumor              -0.056204   0.139460  -0.403 0.686939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2993.1  on 4005  degrees of freedom
## AIC: 3031.1
##
## Number of Fisher Scoring iterations: 5
```

- Our 1st partial test only include the significant predicts on global test, and then we also did a second partial test includes all the covariates that are non significant, and we can see the 1st partial model has an AIC at 3016.9. While the second model with all the non signifcant covraites only get an AIC at 3357.7, which is higher than the first model, so we will definitely not use the model 2 combination. However, we do discover some covariates are significant on the second model like a_stage, and

ln_tumor, so we include these 2 again into the 1st model to get our 3rd model. However, our 3rd model has a higher AIC at 3031.1 than model 1, so we will continue to keep with model 1.

## Step-wise: forward/backward/AiC

### Step-wise for numeric Y

```
numglobal_backward = stepAIC(model_global_num, direction = "backward")
```

```
## Start:  AIC=25070.25
## survival_months ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + age
##
##                       Df Sum of Sq     RSS   AIC
## - differentiate        3     836.3 2023197 25066
## - t_stage              3     846.4 2023207 25066
## - marital_status       4    2945.4 2025306 25068
## - age                  1     576.2 2022937 25069
## - ln_tumor             1     645.1 2023006 25070
## <none>                             2022361 25070
## - progesterone_status  1    1115.3 2023476 25070
## - a_stage              1    1171.0 2023532 25071
## - sqrt_examined        1    1866.6 2024227 25072
## - race                 2    4781.2 2027142 25076
## - estrogen_status      1   13082.5 2035443 25094
## - n_stage              2   20595.8 2042956 25107
##
## Step:  AIC=25065.91
## survival_months ~ race + marital_status + t_stage + n_stage +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + age
##
##                       Df Sum of Sq     RSS   AIC
## - t_stage              3     882.6 2024079 25062
## - marital_status       4    2962.9 2026160 25064
## - age                  1     479.8 2023677 25065
## - ln_tumor             1     716.7 2023914 25065
## <none>                             2023197 25066
## - a_stage              1    1125.8 2024323 25066
## - progesterone_status  1    1248.8 2024446 25066
## - sqrt_examined        1    1874.8 2025072 25068
## - race                 2    5014.2 2028211 25072
## - estrogen_status      1   14398.4 2037595 25092
## - n_stage              2   21605.8 2044803 25105
##
## Step:  AIC=25061.67
## survival_months ~ race + marital_status + n_stage + a_stage +
##     estrogen_status + progesterone_status + ln_tumor + sqrt_examined +
##     age
##
##                       Df Sum of Sq     RSS   AIC
```

```
## - marital_status       4    3040.5 2027120 25060
## - age                  1     490.1 2024570 25061
## <none>                             2024079 25062
## - progesterone_status  1    1249.9 2025329 25062
## - a_stage              1    1279.0 2025358 25062
## - sqrt_examined        1    1802.0 2025881 25063
## - race                 2    4999.6 2029079 25068
## - ln_tumor             1    5102.6 2029182 25070
## - estrogen_status      1   14429.7 2038509 25088
## - n_stage              2   21732.7 2045812 25101
##
## Step:  AIC=25059.71
## survival_months ~ race + n_stage + a_stage + estrogen_status +
##     progesterone_status + ln_tumor + sqrt_examined + age
##
##                       Df Sum of Sq     RSS   AIC
## - age                  1     630.3 2027750 25059
## <none>                             2027120 25060
## - a_stage              1    1348.9 2028469 25060
## - progesterone_status  1    1388.3 2028508 25060
## - sqrt_examined        1    1851.9 2028972 25061
## - race                 2    6008.1 2033128 25068
## - ln_tumor             1    5187.6 2032308 25068
## - estrogen_status      1   14491.8 2041612 25086
## - n_stage              2   22365.2 2049485 25100
##
## Step:  AIC=25058.96
## survival_months ~ race + n_stage + a_stage + estrogen_status +
##     progesterone_status + ln_tumor + sqrt_examined
##
##                       Df Sum of Sq     RSS   AIC
## <none>                             2027750 25059
## - a_stage              1    1315.7 2029066 25060
## - progesterone_status  1    1511.7 2029262 25060
## - sqrt_examined        1    1947.2 2029698 25061
## - ln_tumor             1    4929.4 2032680 25067
## - race                 2    5995.2 2033745 25067
## - estrogen_status      1   14111.6 2041862 25085
## - n_stage              2   22758.1 2050508 25100
```

```
num1_backward = stepAIC(nummodel_partial_1, direction = "backward")
```

```
## Start:  AIC=25075.43
## survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined
##
##                    Df Sum of Sq     RSS   AIC
## <none>                          2037077 25075
## - marital_status    4      4519 2041596 25076
## - sqrt_examined     1      1814 2038891 25077
## - estrogen_status   1     27241 2064317 25127
## - n_stage           2     35253 2072329 25140
```

```r
num2_backward = stepAIC(nummodel_partial_2, direction = "backward")
```

```
## Start:  AIC=25129.5
## survival_months ~ race + t_stage + differentiate + a_stage +
##     progesterone_status + ln_tumor + age
##
##                       Df Sum of Sq     RSS   AIC
## - t_stage              3    1017.9 2061552 25126
## - age                  1     748.5 2061282 25129
## <none>                             2060534 25130
## - ln_tumor             1    1340.1 2061874 25130
## - differentiate        3    3850.4 2064384 25131
## - race                 2    6602.0 2067136 25138
## - a_stage              1    6017.7 2066552 25139
## - progesterone_status  1   13099.9 2073634 25153
##
## Step:  AIC=25125.49
## survival_months ~ race + differentiate + a_stage + progesterone_status +
##     ln_tumor + age
##
##                       Df Sum of Sq     RSS   AIC
## - age                  1     760.6 2062312 25125
## <none>                             2061552 25126
## - differentiate        3    3959.8 2065512 25127
## - race                 2    6616.8 2068169 25134
## - a_stage              1    7095.3 2068647 25137
## - ln_tumor             1   10180.1 2071732 25143
## - progesterone_status  1   13093.7 2074645 25149
##
## Step:  AIC=25124.97
## survival_months ~ race + differentiate + a_stage + progesterone_status +
##     ln_tumor
##
##                       Df Sum of Sq     RSS   AIC
## <none>                             2062312 25125
## - differentiate        3    3693.9 2066006 25126
## - race                 2    6605.5 2068918 25134
## - a_stage              1    7048.3 2069361 25137
## - ln_tumor             1    9857.2 2072170 25142
## - progesterone_status  1   13410.4 2075723 25149
```

```r
num3_backward = stepAIC(nummodel_partial_3, direction = "backward")
```

```
## Start:  AIC=25066.58
## survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined + a_stage + progesterone_status + ln_tumor
##
##                       Df Sum of Sq     RSS   AIC
## <none>                             2029574 25067
## - marital_status       4    4171.9 2033745 25067
## - a_stage              1    1246.9 2030820 25067
## - progesterone_status  1    1383.5 2030957 25067
```

```
## - sqrt_examined        1    1942.7 2031516 25068
## - ln_tumor             1    4544.3 2034118 25074
## - estrogen_status      1   14527.3 2044101 25093
## - n_stage              2   22634.1 2052208 25107
```

```
numglobal_backward = stepAIC(model_global_num, direction = "forward")
```

```
## Start:  AIC=25070.25
## survival_months ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + age
```

```
num1_backward = stepAIC(nummodel_partial_1, direction = "forward")
```

```
## Start:  AIC=25075.43
## survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined
```

```
num2_backward = stepAIC(nummodel_partial_2, direction = "forward")
```

```
## Start:  AIC=25129.5
## survival_months ~ race + t_stage + differentiate + a_stage +
##     progesterone_status + ln_tumor + age
```

```
num3_backward = stepAIC(nummodel_partial_3, direction = "forward")
```

```
## Start:  AIC=25066.58
## survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined + a_stage + progesterone_status + ln_tumor
```

```
numglobal_backward = stepAIC(model_global_num, direction = "both")
```

```
## Start:  AIC=25070.25
## survival_months ~ race + marital_status + t_stage + n_stage +
##     differentiate + a_stage + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + age
##
##                       Df Sum of Sq     RSS   AIC
## - differentiate        3     836.3 2023197 25066
## - t_stage              3     846.4 2023207 25066
## - marital_status       4    2945.4 2025306 25068
## - age                  1     576.2 2022937 25069
## - ln_tumor             1     645.1 2023006 25070
## <none>                             2022361 25070
## - progesterone_status  1    1115.3 2023476 25070
## - a_stage              1    1171.0 2023532 25071
## - sqrt_examined        1    1866.6 2024227 25072
## - race                 2    4781.2 2027142 25076
## - estrogen_status      1   13082.5 2035443 25094
## - n_stage              2   20595.8 2042956 25107
```

```
##
## Step:  AIC=25065.91
## survival_months ~ race + marital_status + t_stage + n_stage +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + age
##
##                       Df Sum of Sq     RSS   AIC
## - t_stage              3     882.6 2024079 25062
## - marital_status       4    2962.9 2026160 25064
## - age                  1     479.8 2023677 25065
## - ln_tumor             1     716.7 2023914 25065
## <none>                             2023197 25066
## - a_stage              1    1125.8 2024323 25066
## - progesterone_status  1    1248.8 2024446 25066
## - sqrt_examined        1    1874.8 2025072 25068
## + differentiate        3     836.3 2022361 25070
## - race                 2    5014.2 2028211 25072
## - estrogen_status      1   14398.4 2037595 25092
## - n_stage              2   21605.8 2044803 25105
##
## Step:  AIC=25061.67
## survival_months ~ race + marital_status + n_stage + a_stage +
##     estrogen_status + progesterone_status + ln_tumor + sqrt_examined +
##     age
##
##                       Df Sum of Sq     RSS   AIC
## - marital_status       4    3040.5 2027120 25060
## - age                  1     490.1 2024570 25061
## <none>                             2024079 25062
## - progesterone_status  1    1249.9 2025329 25062
## - a_stage              1    1279.0 2025358 25062
## - sqrt_examined        1    1802.0 2025881 25063
## + t_stage              3     882.6 2023197 25066
## + differentiate        3     872.5 2023207 25066
## - race                 2    4999.6 2029079 25068
## - ln_tumor             1    5102.6 2029182 25070
## - estrogen_status      1   14429.7 2038509 25088
## - n_stage              2   21732.7 2045812 25101
##
## Step:  AIC=25059.71
## survival_months ~ race + n_stage + a_stage + estrogen_status +
##     progesterone_status + ln_tumor + sqrt_examined + age
##
##                       Df Sum of Sq     RSS   AIC
## - age                  1     630.3 2027750 25059
## <none>                             2027120 25060
## - a_stage              1    1348.9 2028469 25060
## - progesterone_status  1    1388.3 2028508 25060
## - sqrt_examined        1    1851.9 2028972 25061
## + marital_status       4    3040.5 2024079 25062
## + t_stage              3     960.2 2026160 25064
## + differentiate        3     889.9 2026230 25064
## - race                 2    6008.1 2033128 25068
## - ln_tumor             1    5187.6 2032308 25068
```

```
## - estrogen_status      1   14491.8 2041612 25086
## - n_stage              2   22365.2 2049485 25100
##
## Step:  AIC=25058.96
## survival_months ~ race + n_stage + a_stage + estrogen_status +
##     progesterone_status + ln_tumor + sqrt_examined
##
##                        Df Sum of Sq     RSS   AIC
## <none>                              2027750 25059
## - a_stage               1    1315.7 2029066 25060
## + age                   1     630.3 2027120 25060
## - progesterone_status   1    1511.7 2029262 25060
## + marital_status        4    3180.7 2024570 25061
## - sqrt_examined         1    1947.2 2029698 25061
## + t_stage               3     972.1 2026778 25063
## + differentiate         3     779.2 2026971 25063
## - ln_tumor              1    4929.4 2032680 25067
## - race                  2    5995.2 2033745 25067
## - estrogen_status       1   14111.6 2041862 25085
## - n_stage               2   22758.1 2050508 25100
```

```
num1_backward = stepAIC(nummodel_partial_1, direction = "both")
```

```
## Start:  AIC=25075.43
## survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined
##
##                    Df Sum of Sq     RSS   AIC
## <none>                          2037077 25075
## - marital_status    4      4519 2041596 25076
## - sqrt_examined     1      1814 2038891 25077
## - estrogen_status   1     27241 2064317 25127
## - n_stage           2     35253 2072329 25140
```

```
num2_backward = stepAIC(nummodel_partial_2, direction = "both")
```

```
## Start:  AIC=25129.5
## survival_months ~ race + t_stage + differentiate + a_stage +
##     progesterone_status + ln_tumor + age
##
##                        Df Sum of Sq     RSS   AIC
## - t_stage               3    1017.9 2061552 25126
## - age                   1     748.5 2061282 25129
## <none>                              2060534 25130
## - ln_tumor              1    1340.1 2061874 25130
## - differentiate         3    3850.4 2064384 25131
## - race                  2    6602.0 2067136 25138
## - a_stage               1    6017.7 2066552 25139
## - progesterone_status   1   13099.9 2073634 25153
##
## Step:  AIC=25125.49
## survival_months ~ race + differentiate + a_stage + progesterone_status +
##     ln_tumor + age
```

```
##
##                        Df Sum of Sq     RSS   AIC
## - age                   1     760.6 2062312 25125
## <none>                              2061552 25126
## - differentiate         3    3959.8 2065512 25127
## + t_stage               3    1017.9 2060534 25130
## - race                  2    6616.8 2068169 25134
## - a_stage               1    7095.3 2068647 25137
## - ln_tumor              1   10180.1 2071732 25143
## - progesterone_status   1   13093.7 2074645 25149
##
## Step:  AIC=25124.97
## survival_months ~ race + differentiate + a_stage + progesterone_status +
##     ln_tumor
##
##                        Df Sum of Sq     RSS   AIC
## <none>                              2062312 25125
## + age                   1     760.6 2061552 25126
## - differentiate         3    3693.9 2066006 25126
## + t_stage               3    1030.1 2061282 25129
## - race                  2    6605.5 2068918 25134
## - a_stage               1    7048.3 2069361 25137
## - ln_tumor              1    9857.2 2072170 25142
## - progesterone_status   1   13410.4 2075723 25149
```

```
num3_backward = stepAIC(nummodel_partial_3, direction = "both")
```

```
## Start:  AIC=25066.58
## survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined + a_stage + progesterone_status + ln_tumor
##
##                        Df Sum of Sq     RSS   AIC
## <none>                              2029574 25067
## - marital_status        4    4171.9 2033745 25067
## - a_stage               1    1246.9 2030820 25067
## - progesterone_status   1    1383.5 2030957 25067
## - sqrt_examined         1    1942.7 2031516 25068
## - ln_tumor              1    4544.3 2034118 25074
## - estrogen_status       1   14527.3 2044101 25093
## - n_stage               2   22634.1 2052208 25107
```

- Above all, we can see that from all the backward, forward, and stepwise selection, we have the same AIC get from the 4 numerical model, and the best combination with the lowest AIC of 25062.62 is the combination of progesterone_status, a_stage, sqrt_examined, marital_status, ln_tumor, estrogen_status n_stage

### Step-wise for binary Y

```
binglobal_backward = stepAIC(model_global_bin, direction = "backward")
```

```
## Start:  AIC=3020.64
```

```
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + age
##
##                        Df Deviance    AIC
## - a_stage               1   2978.6 3018.6
## - ln_tumor              1   2978.9 3018.9
## <none>                      2978.6 3020.6
## - marital_status        4   2987.5 3021.5
## - t_stage               3   2990.2 3026.2
## - race                  2   2993.1 3031.1
## - sqrt_examined         1   2993.7 3033.7
## - estrogen_status       1   2995.5 3035.5
## - progesterone_status   1   2997.9 3037.9
## - age                   1   2998.3 3038.3
## - differentiate         3   3011.0 3047.0
## - n_stage               2   3107.3 3145.3
##
## Step:  AIC=3018.64
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + ln_tumor + sqrt_examined +
##     age
##
##                        Df Deviance    AIC
## - ln_tumor              1   2978.9 3016.9
## <none>                      2978.6 3018.6
## - marital_status        4   2987.5 3019.5
## - t_stage               3   2991.2 3025.2
## - race                  2   2993.1 3029.1
## - sqrt_examined         1   2993.7 3031.7
## - estrogen_status       1   2995.5 3033.5
## - progesterone_status   1   2997.9 3035.9
## - age                   1   2998.3 3036.3
## - differentiate         3   3011.0 3045.0
## - n_stage               2   3113.8 3149.8
##
## Step:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                        Df Deviance    AIC
## <none>                      2978.9 3016.9
## - marital_status        4   2987.8 3017.8
## - race                  2   2993.2 3027.2
## - sqrt_examined         1   2994.0 3030.0
## - estrogen_status       1   2995.7 3031.7
## - progesterone_status   1   2998.1 3034.1
## - age                   1   2998.4 3034.4
## - t_stage               3   3008.6 3040.6
## - differentiate         3   3011.6 3043.6
## - n_stage               2   3116.4 3150.4
```

```r
bin1_backward = stepAIC(binmodel_partial_1, direction = "backward")
```

```
## Start:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                       Df Deviance    AIC
## <none>                     2978.9 3016.9
## - marital_status       4   2987.8 3017.8
## - race                 2   2993.2 3027.2
## - sqrt_examined        1   2994.0 3030.0
## - estrogen_status      1   2995.7 3031.7
## - progesterone_status  1   2998.1 3034.1
## - age                  1   2998.4 3034.4
## - t_stage              3   3008.6 3040.6
## - differentiate        3   3011.6 3043.6
## - n_stage              2   3116.4 3150.4
```

```
bin2_backward = stepAIC(binmodel_partial_2, direction = "backward")
```

```
## Start:  AIC=3357.74
## status ~ a_stage + ln_tumor
##
##            Df Deviance    AIC
## <none>          3351.7 3357.7
## - a_stage   1   3370.9 3374.9
## - ln_tumor  1   3415.7 3419.7
```

```
bin3_backward = stepAIC(binmodel_partial_3, direction = "backward")
```

```
## Start:  AIC=3031.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     a_stage + ln_tumor
##
##                       Df Deviance    AIC
## - a_stage              1   2993.1 3029.1
## - ln_tumor             1   2993.2 3029.2
## <none>                     2993.1 3031.1
## - marital_status       4   3005.1 3035.1
## - t_stage              3   3004.8 3036.8
## - sqrt_examined        1   3008.6 3044.6
## - estrogen_status      1   3010.1 3046.1
## - progesterone_status  1   3012.4 3048.4
## - age                  1   3013.0 3049.0
## - differentiate        3   3027.6 3059.6
## - n_stage              2   3123.8 3157.8
##
## Step:  AIC=3029.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     ln_tumor
##
##                       Df Deviance    AIC
## - ln_tumor             1   2993.2 3027.2
```

```
## <none>                     2993.1 3029.1
## - marital_status       4   3005.1 3033.1
## - t_stage              3   3005.8 3035.8
## - sqrt_examined        1   3008.6 3042.6
## - estrogen_status      1   3010.1 3044.1
## - progesterone_status  1   3012.4 3046.4
## - age                  1   3013.1 3047.1
## - differentiate        3   3027.6 3057.6
## - n_stage              2   3130.2 3162.2
##
## Step:  AIC=3027.25
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                       Df Deviance    AIC
## <none>                     2993.2 3027.3
## - marital_status       4   3005.3 3031.3
## - sqrt_examined        1   3008.8 3040.8
## - estrogen_status      1   3010.3 3042.3
## - progesterone_status  1   3012.6 3044.6
## - age                  1   3013.1 3045.1
## - t_stage              3   3022.2 3050.2
## - differentiate        3   3028.1 3056.1
## - n_stage              2   3132.5 3162.5
```

From the backward elimination, the best model with lowest AIC of 3016.87 is: status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + progesterone_status + sqrt_examined + age.

```
binglobal_backward = stepAIC(model_global_bin, direction = "forward")
```

```
## Start:  AIC=3020.64
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + age
```

```
bin1_backward = stepAIC(binmodel_partial_1, direction = "forward")
```

```
## Start:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
```

```
bin2_backward = stepAIC(binmodel_partial_2, direction = "forward")
```

```
## Start:  AIC=3357.74
## status ~ a_stage + ln_tumor
```

```
bin3_backward = stepAIC(binmodel_partial_3, direction = "forward")
```

```
## Start:  AIC=3031.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     a_stage + ln_tumor
```

From the forward selection, the best model with lowest AIC of 3016.87 status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + progesterone_status + sqrt_examined + age. Backward and forward elimination selection both produce same AIC value and the models are the same.

```
binglobal_backward = stepAIC(model_global_bin, direction = "both")
```

```
## Start:  AIC=3020.64
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + age
##
##                        Df Deviance    AIC
## - a_stage               1   2978.6 3018.6
## - ln_tumor              1   2978.9 3018.9
## <none>                      2978.6 3020.6
## - marital_status        4   2987.5 3021.5
## - t_stage               3   2990.2 3026.2
## - race                  2   2993.1 3031.1
## - sqrt_examined         1   2993.7 3033.7
## - estrogen_status       1   2995.5 3035.5
## - progesterone_status   1   2997.9 3037.9
## - age                   1   2998.3 3038.3
## - differentiate         3   3011.0 3047.0
## - n_stage               2   3107.3 3145.3
##
## Step:  AIC=3018.64
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + ln_tumor + sqrt_examined +
##     age
##
##                        Df Deviance    AIC
## - ln_tumor              1   2978.9 3016.9
## <none>                      2978.6 3018.6
## - marital_status        4   2987.5 3019.5
## + a_stage               1   2978.6 3020.6
## - t_stage               3   2991.2 3025.2
## - race                  2   2993.1 3029.1
## - sqrt_examined         1   2993.7 3031.7
## - estrogen_status       1   2995.5 3033.5
## - progesterone_status   1   2997.9 3035.9
## - age                   1   2998.3 3036.3
## - differentiate         3   3011.0 3045.0
## - n_stage               2   3113.8 3149.8
##
## Step:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                        Df Deviance    AIC
## <none>                      2978.9 3016.9
## - marital_status        4   2987.8 3017.8
## + ln_tumor              1   2978.6 3018.6
## + a_stage               1   2978.9 3018.9
## - race                  2   2993.2 3027.2
```

```
## - sqrt_examined        1   2994.0 3030.0
## - estrogen_status      1   2995.7 3031.7
## - progesterone_status  1   2998.1 3034.1
## - age                  1   2998.4 3034.4
## - t_stage              3   3008.6 3040.6
## - differentiate        3   3011.6 3043.6
## - n_stage              2   3116.4 3150.4
```

```
bin1_backward = stepAIC(binmodel_partial_1, direction = "both")
```

```
## Start:  AIC=3016.87
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                       Df Deviance    AIC
## <none>                     2978.9 3016.9
## - marital_status       4   2987.8 3017.8
## - race                 2   2993.2 3027.2
## - sqrt_examined        1   2994.0 3030.0
## - estrogen_status      1   2995.7 3031.7
## - progesterone_status  1   2998.1 3034.1
## - age                  1   2998.4 3034.4
## - t_stage              3   3008.6 3040.6
## - differentiate        3   3011.6 3043.6
## - n_stage              2   3116.4 3150.4
```

```
bin2_backward = stepAIC(binmodel_partial_2, direction = "both")
```

```
## Start:  AIC=3357.74
## status ~ a_stage + ln_tumor
##
##            Df Deviance    AIC
## <none>          3351.7 3357.7
## - a_stage   1   3370.9 3374.9
## - ln_tumor  1   3415.7 3419.7
```

```
bin3_backward = stepAIC(binmodel_partial_3, direction = "both")
```

```
## Start:  AIC=3031.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     a_stage + ln_tumor
##
##                       Df Deviance    AIC
## - a_stage              1   2993.1 3029.1
## - ln_tumor             1   2993.2 3029.2
## <none>                     2993.1 3031.1
## - marital_status       4   3005.1 3035.1
## - t_stage              3   3004.8 3036.8
## - sqrt_examined        1   3008.6 3044.6
## - estrogen_status      1   3010.1 3046.1
## - progesterone_status  1   3012.4 3048.4
```

```
## - age                  1   3013.0 3049.0
## - differentiate        3   3027.6 3059.6
## - n_stage              2   3123.8 3157.8
##
## Step:  AIC=3029.09
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age +
##     ln_tumor
##
##                       Df Deviance    AIC
## - ln_tumor             1   2993.2 3027.2
## <none>                     2993.1 3029.1
## + a_stage              1   2993.1 3031.1
## - marital_status       4   3005.1 3033.1
## - t_stage              3   3005.8 3035.8
## - sqrt_examined        1   3008.6 3042.6
## - estrogen_status      1   3010.1 3044.1
## - progesterone_status  1   3012.4 3046.4
## - age                  1   3013.1 3047.1
## - differentiate        3   3027.6 3057.6
## - n_stage              2   3130.2 3162.2
##
## Step:  AIC=3027.25
## status ~ marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##
##                       Df Deviance    AIC
## <none>                     2993.2 3027.3
## + ln_tumor             1   2993.1 3029.1
## + a_stage              1   2993.2 3029.3
## - marital_status       4   3005.3 3031.3
## - sqrt_examined        1   3008.8 3040.8
## - estrogen_status      1   3010.3 3042.3
## - progesterone_status  1   3012.6 3044.6
## - age                  1   3013.1 3045.1
## - t_stage              3   3022.2 3050.2
## - differentiate        3   3028.1 3056.1
## - n_stage              2   3132.5 3162.5
```

- From the stepwise regression, we have the lowest AIC of 3016.98 which is the model of status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + progesterone_status + sqrt_examined + age. The model is the same compared to the one we picked from the backward elimination and forward elimination. We will go with this model for further statistical testing.

## ANOVA

```r
# anova for numeric Y
newbc1$ln_tumor <- ifelse(newbc1$ln_tumor <= 0, NA, newbc1$ln_tumor)

# Apply transformations
newbc1$sqrt_examined <- sqrt(newbc1$sqrt_examined)
```

```r
newbc1$ln_tumor <- log(newbc1$ln_tumor)

# If you decide to remove rows with any NA values
newbc1 <- na.omit(newbc1)

# Fit the linear model
lm_model <- lm(survival_months ~ marital_status + n_stage + estrogen_status +
               sqrt_examined + a_stage + progesterone_status + ln_tumor,
               data = newbc1)

# Perform ANOVA
anova_result <- anova(lm_model)
print(anova_result)
```

**anova for numeric Y**

```
## Analysis of Variance Table
##
## Response: survival_months
##                       Df  Sum Sq Mean Sq F value    Pr(>F)
## marital_status         4    6820  1705.1  3.3664  0.009289 **
## n_stage                2   40939 20469.4 40.4129 < 2.2e-16 ***
## estrogen_status        1   27053 27053.0 53.4111 3.249e-13 ***
## sqrt_examined          1    1667  1667.4  3.2919  0.069695 .
## a_stage                1    1407  1407.4  2.7786  0.095607 .
## progesterone_status    1    1590  1590.4  3.1400  0.076472 .
## ln_tumor               1    3395  3395.3  6.7033  0.009658 **
## Residuals           4005 2028556   506.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on ANOVA test, matial_status, n_stage, estrogen_status, and ln_tumors are significant variables.

```r
# delete insignificant variables from the above anova table and fit a new model
lm_model1 <- lm(survival_months ~ marital_status + n_stage + estrogen_status + ln_tumor,
               data = newbc1)

# Perform ANOVA
anova_result1 <- anova(lm_model1)
print(anova_result1)
```

```
## Analysis of Variance Table
##
## Response: survival_months
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## marital_status     4    6820  1705.1  3.3611  0.009374 **
## n_stage            2   40939 20469.4 40.3500 < 2.2e-16 ***
## estrogen_status    1   27053 27053.0 53.3279 3.387e-13 ***
## ln_tumor           1    3375  3375.0  6.6529  0.009935 **
## Residuals       4008 2033241   507.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Partial f test for nested model
model_reduced <- lm(survival_months ~ marital_status + n_stage + estrogen_status + ln_tumor,
                    data = newbc1)
model_full <- lm(survival_months ~ marital_status + n_stage + estrogen_status +
                 sqrt_examined + a_stage + progesterone_status + ln_tumor,
                 data = newbc1)
anova(model_reduced, model_full)
```

```
## Analysis of Variance Table
##
## Model 1: survival_months ~ marital_status + n_stage + estrogen_status +
##     ln_tumor
## Model 2: survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined + a_stage + progesterone_status + ln_tumor
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1   4008 2033241
## 2   4005 2028556  3    4685.5 3.0835 0.02625 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With p significant p value, we have enough evidence to show that the full model is better.

```r
model2 <- glm(status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + pr
anova_result1 <- anova(model2)
print(anova_result1)
```

**anova for binary Y**

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: status
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev
## NULL                                 4023     521.70
## race                 2   3.6262      4021     518.08
## marital_status       4   2.9196      4017     515.16
## t_stage              3  13.1374      4014     502.02
## n_stage              2  25.1703      4012     476.85
## differentiate        3   6.8813      4009     469.97
## estrogen_status      1   8.9916      4008     460.98
## progesterone_status  1   2.9595      4007     458.02
## sqrt_examined        1   1.8003      4006     456.22
## age                  1   2.1022      4005     454.11
```

```r
summary(model2)
```

```
##
## Call:
## glm(formula = status ~ race + marital_status + t_stage + n_stage +
##     differentiate + estrogen_status + progesterone_status + sqrt_examined +
##     age, data = newbc2)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.7708680  0.0439145  17.554  < 2e-16 ***
## race2                  -0.0681853  0.0209930  -3.248 0.001172 **
## race3                   0.0416639  0.0198151   2.103 0.035560 *
## marital_status2        -0.0259111  0.0167149  -1.550 0.121178
## marital_status3        -0.0175357  0.0153627  -1.141 0.253752
## marital_status4        -0.0321326  0.0235225  -1.366 0.172004
## marital_status5        -0.1345247  0.0508574  -2.645 0.008198 **
## t_stage2               -0.0396395  0.0118704  -3.339 0.000847 ***
## t_stage3               -0.0603490  0.0174234  -3.464 0.000538 ***
## t_stage4               -0.1748054  0.0350531  -4.987 6.40e-07 ***
## n_stage2               -0.0837902  0.0139153  -6.021 1.88e-09 ***
## n_stage3               -0.2421576  0.0183695 -13.183  < 2e-16 ***
## differentiate2          0.0509827  0.0127519   3.998 6.50e-05 ***
## differentiate3          0.0895198  0.0183190   4.887 1.07e-06 ***
## differentiate4         -0.1824586  0.0781803  -2.334 0.019654 *
## estrogen_status1        0.1408815  0.0252403   5.582 2.54e-08 ***
## progesterone_status1    0.0779114  0.0164916   4.724 2.39e-06 ***
## sqrt_examined           0.0191703  0.0050357   3.807 0.000143 ***
## age                    -0.0026658  0.0006191  -4.306 1.70e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1133866)
##
##     Null deviance: 521.70  on 4023  degrees of freedom
## Residual deviance: 454.11  on 4005  degrees of freedom
## AIC: 2680.5
##
## Number of Fisher Scoring iterations: 2
```

```r
null_model <- glm(status ~ 1, family = binomial, data = newbc2)

anova(null_model, model2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: status ~ 1
## Model 2: status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      4023     3444.7
## 2      4005      454.1 18   2990.6 < 2.2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With p significant p value, we have enough evidence to show that the full model is better, which is status ~ race + marital_status + t_stage + n_stage + differentiate + estrogen_status + progesterone_status + sqrt_examined + age

## Criterion procedures

```
# AIC-based selection for the numeric outcome
aic_selected_num_model <- stepAIC(lm_model, direction = "both")
```

```
## Start:  AIC=25027.99
## survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined + a_stage + progesterone_status + ln_tumor
##
##                       Df Sum of Sq     RSS   AIC
## <none>                              2028556 25028
## - marital_status       4    4192.1 2032748 25028
## - a_stage              1    1306.7 2029862 25029
## - progesterone_status  1    1506.9 2030063 25029
## - sqrt_examined        1    1812.4 2030368 25030
## - ln_tumor             1    3395.3 2031951 25033
## - estrogen_status      1   14469.8 2043025 25054
## - n_stage              2   23667.3 2052223 25071
```

```
summary(aic_selected_num_model)
```

```
##
## Call:
## lm(formula = survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined + a_stage + progesterone_status + ln_tumor,
##     data = newbc1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.517 -15.640   1.105  18.394  52.990
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           60.2697     3.9702  15.181  < 2e-16 ***
## marital_status2       -1.0335     1.1130  -0.929  0.35315
## marital_status3       -1.2632     1.0090  -1.252  0.21066
## marital_status4       -2.1402     1.5343  -1.395  0.16311
## marital_status5       -7.8496     3.3909  -2.315  0.02067 *
## n_stage2              -2.9220     0.9277  -3.150  0.00165 **
## n_stage3              -8.3923     1.2582  -6.670 2.91e-11 ***
## estrogen_status1       8.8846     1.6623   5.345 9.55e-08 ***
## sqrt_examined          2.2560     1.1926   1.892  0.05861 .
## a_stage1               4.0047     2.4934   1.606  0.10832
## progesterone_status1   1.8906     1.0961   1.725  0.08463 .
```

```
## ln_tumor              -4.2533     1.6428  -2.589  0.00966 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.51 on 4005 degrees of freedom
## Multiple R-squared:  0.03925,    Adjusted R-squared:  0.03661
## F-statistic: 14.87 on 11 and 4005 DF,  p-value: < 2.2e-16
```

```
# BIC-based selection (which is similar to AIC but with a larger penalty for the number of parameters)
bic_selected_num_model <- stepAIC(lm_model, direction = "both", k = log(nrow(newbc1)))
```

```
## Start:  AIC=25103.57
## survival_months ~ marital_status + n_stage + estrogen_status +
##     sqrt_examined + a_stage + progesterone_status + ln_tumor
##
##                        Df Sum of Sq     RSS   AIC
## - marital_status        4    4192.1 2032748 25079
## - a_stage               1    1306.7 2029862 25098
## - progesterone_status   1    1506.9 2030063 25098
## - sqrt_examined         1    1812.4 2030368 25099
## - ln_tumor              1    3395.3 2031951 25102
## <none>                              2028556 25104
## - estrogen_status       1   14469.8 2043025 25124
## - n_stage               2   23667.3 2052223 25134
##
## Step:  AIC=25078.67
## survival_months ~ n_stage + estrogen_status + sqrt_examined +
##     a_stage + progesterone_status + ln_tumor
##
##                        Df Sum of Sq     RSS   AIC
## - a_stage               1    1377.5 2034125 25073
## - progesterone_status   1    1701.5 2034449 25074
## - sqrt_examined         1    1910.2 2034658 25074
## - ln_tumor              1    3399.5 2036147 25077
## <none>                              2032748 25079
## - estrogen_status       1   14532.1 2047280 25099
## + marital_status        4    4192.1 2028556 25104
## - n_stage               2   24643.1 2057391 25110
##
## Step:  AIC=25073.1
## survival_months ~ n_stage + estrogen_status + sqrt_examined +
##     progesterone_status + ln_tumor
##
##                        Df Sum of Sq     RSS   AIC
## - progesterone_status   1    1626.5 2035752 25068
## - sqrt_examined         1    1987.6 2036113 25069
## - ln_tumor              1    3579.3 2037705 25072
## <none>                              2034125 25073
## + a_stage               1    1377.5 2032748 25079
## - estrogen_status       1   14980.2 2049105 25094
## + marital_status        4    4263.0 2029862 25098
## - n_stage               2   29815.4 2063941 25115
##
## Step:  AIC=25068.01
```

```
## survival_months ~ n_stage + estrogen_status + sqrt_examined +
##     ln_tumor
##
##                        Df Sum of Sq     RSS   AIC
## - sqrt_examined         1    2050.1 2037802 25064
## - ln_tumor              1    3662.4 2039414 25067
## <none>                               2035752 25068
## + progesterone_status   1    1626.5 2034125 25073
## + a_stage               1    1302.5 2034449 25074
## + marital_status        4    4451.5 2031300 25092
## - n_stage               2   30573.5 2066325 25111
## - estrogen_status       1   27565.1 2063317 25114
##
## Step:  AIC=25063.75
## survival_months ~ n_stage + estrogen_status + ln_tumor
##
##                        Df Sum of Sq     RSS   AIC
## - ln_tumor              1    3380.8 2041183 25062
## <none>                               2037802 25064
## + sqrt_examined         1    2050.1 2035752 25068
## + progesterone_status   1    1689.0 2036113 25069
## + a_stage               1    1377.6 2036424 25069
## + marital_status        4    4560.8 2033241 25088
## - n_stage               2   28586.3 2066388 25103
## - estrogen_status       1   27469.9 2065272 25109
##
## Step:  AIC=25062.11
## survival_months ~ n_stage + estrogen_status
##
##                        Df Sum of Sq     RSS   AIC
## <none>                               2041183 25062
## + ln_tumor              1      3381 2037802 25064
## + sqrt_examined         1      1768 2039414 25067
## + progesterone_status   1      1766 2039417 25067
## + a_stage               1      1545 2039638 25067
## + marital_status        4      4567 2036616 25086
## - estrogen_status       1     27657 2068840 25108
## - n_stage               2     35418 2076601 25115
```

```
summary(bic_selected_num_model)
```

```
##
## Call:
## lm(formula = survival_months ~ n_stage + estrogen_status, data = newbc1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.709 -15.709   1.291  18.423  53.008
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       63.1554     1.4242  44.344  < 2e-16 ***
## n_stage2          -3.1318     0.8993  -3.483 0.000502 ***
## n_stage3          -9.1636     1.1309  -8.103 7.05e-16 ***
```

```
## estrogen_status1  10.5533     1.4312   7.374 2.00e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.55 on 4013 degrees of freedom
## Multiple R-squared:  0.03327,    Adjusted R-squared:  0.03255
## F-statistic: 46.03 on 3 and 4013 DF,  p-value: < 2.2e-16
```

According to the result, we can see that based on the AIC selection, the final model according to the survival_month is survival_months ~ marital_status + n_stage + estrogen_status + sqrt_examined + a_stage + progesterone_status + ln_tumor.

According to the BIC selection of the final result, the final model they selected is survival_months ~ n_stage + estrogen_status.

```
# For the binary outcome, we will use the AIC criterion as well
aic_selected_bin_model <- stepAIC(glmfit, direction = "both")
```

```
## Start:  AIC=2997.39
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + reginol_node_positive + age
##
##                          Df Deviance    AIC
## - a_stage                 1   2953.4 2995.4
## - ln_tumor                1   2953.6 2995.6
## - marital_status          4   2961.2 2997.2
## <none>                        2953.4 2997.4
## - t_stage                 3   2964.7 3002.7
## - n_stage                 2   2967.3 3007.3
## - race                    2   2969.2 3009.2
## - estrogen_status         1   2970.5 3012.5
## - age                     1   2971.4 3013.4
## - progesterone_status     1   2973.6 3015.6
## - reginol_node_positive   1   2978.6 3020.6
## - sqrt_examined           1   2980.4 3022.4
## - differentiate           3   2986.8 3024.8
##
## Step:  AIC=2995.41
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + ln_tumor + sqrt_examined +
##     reginol_node_positive + age
##
##                          Df Deviance    AIC
## - ln_tumor                1   2953.6 2993.6
## - marital_status          4   2961.2 2995.2
## <none>                        2953.4 2995.4
## + a_stage                 1   2953.4 2997.4
## - t_stage                 3   2966.0 3002.0
## - n_stage                 2   2967.5 3005.5
## - race                    2   2969.2 3007.2
## - estrogen_status         1   2970.6 3010.6
## - age                     1   2971.4 3011.4
## - progesterone_status     1   2973.6 3013.6
```

```
## - reginol_node_positive  1   2978.6 3018.6
## - sqrt_examined          1   2980.4 3020.4
## - differentiate          3   2986.8 3022.8
##
## Step:  AIC=2993.57
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + reginol_node_positive +
##     age
##
##                         Df Deviance    AIC
## - marital_status         4   2961.5 2993.5
## <none>                       2953.6 2993.6
## + ln_tumor               1   2953.4 2995.4
## + a_stage                1   2953.6 2995.6
## - n_stage                2   2967.9 3003.9
## - race                   2   2969.3 3005.3
## - estrogen_status        1   2970.7 3008.7
## - age                    1   2971.4 3009.4
## - progesterone_status    1   2973.8 3011.8
## - reginol_node_positive  1   2978.9 3016.9
## - t_stage                3   2983.2 3017.2
## - sqrt_examined          1   2980.6 3018.6
## - differentiate          3   2987.3 3021.3
##
## Step:  AIC=2993.48
## status ~ race + t_stage + n_stage + differentiate + estrogen_status +
##     progesterone_status + sqrt_examined + reginol_node_positive +
##     age
##
##                         Df Deviance    AIC
## <none>                       2961.5 2993.5
## + marital_status         4   2953.6 2993.6
## + ln_tumor               1   2961.2 2995.2
## + a_stage                1   2961.4 2995.4
## - n_stage                2   2975.7 3003.7
## - race                   2   2980.3 3008.3
## - estrogen_status        1   2978.4 3008.4
## - age                    1   2981.4 3011.4
## - progesterone_status    1   2982.6 3012.6
## - reginol_node_positive  1   2987.8 3017.8
## - t_stage                3   2992.1 3018.1
## - sqrt_examined          1   2989.1 3019.1
## - differentiate          3   2995.1 3021.1
```

```
summary(aic_selected_bin_model)
```

```
##
## Call:
## glm(formula = status ~ race + t_stage + n_stage + differentiate +
##     estrogen_status + progesterone_status + sqrt_examined + reginol_node_positive +
##     age, family = binomial, data = newbc2)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)            1.623139   0.371232   4.372 1.23e-05 ***
## race2                 -0.574227   0.158413  -3.625 0.000289 ***
## race3                  0.430601   0.201848   2.133 0.032901 *
## t_stage2              -0.429230   0.112954  -3.800 0.000145 ***
## t_stage3              -0.561433   0.148563  -3.779 0.000157 ***
## t_stage4              -1.122611   0.242936  -4.621 3.82e-06 ***
## n_stage2              -0.474258   0.128697  -3.685 0.000229 ***
## n_stage3              -0.644148   0.234160  -2.751 0.005943 **
## differentiate2         0.389412   0.104623   3.722 0.000198 ***
## differentiate3         0.908192   0.191600   4.740 2.14e-06 ***
## differentiate4        -0.954165   0.525800  -1.815 0.069571 .
## estrogen_status1       0.731594   0.176812   4.138 3.51e-05 ***
## progesterone_status1   0.600531   0.127305   4.717 2.39e-06 ***
## sqrt_examined          0.259110   0.049645   5.219 1.80e-07 ***
## reginol_node_positive -0.076301   0.014939  -5.107 3.27e-07 ***
## age                   -0.024052   0.005443  -4.419 9.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2961.5  on 4008  degrees of freedom
## AIC: 2993.5
##
## Number of Fisher Scoring iterations: 5
```

```
# And the BIC criterion
bic_selected_bin_model <- stepAIC(glmfit, direction = "both", family = binomial, k = log(nrow(newbc2)))
```

```
## Start:  AIC=3135.99
## status ~ race + marital_status + t_stage + n_stage + differentiate +
##     a_stage + estrogen_status + progesterone_status + ln_tumor +
##     sqrt_examined + reginol_node_positive + age
##
##                         Df Deviance    AIC
## - marital_status         4   2961.2 3110.6
## - t_stage                3   2964.7 3122.4
## - a_stage                1   2953.4 3127.7
## - ln_tumor               1   2953.6 3127.8
## - n_stage                2   2967.3 3133.3
## - race                   2   2969.2 3135.2
## <none>                       2953.4 3136.0
## - differentiate          3   2986.8 3144.5
## - estrogen_status        1   2970.5 3144.8
## - age                    1   2971.4 3145.7
## - progesterone_status    1   2973.6 3147.9
## - reginol_node_positive  1   2978.6 3152.9
## - sqrt_examined          1   2980.4 3154.7
##
## Step:  AIC=3110.59
## status ~ race + t_stage + n_stage + differentiate + a_stage +
##     estrogen_status + progesterone_status + ln_tumor + sqrt_examined +
##     reginol_node_positive + age
```

```
##
##                          Df Deviance    AIC
## - t_stage                 3   2972.7 3097.2
## - a_stage                 1   2961.2 3102.3
## - ln_tumor                1   2961.4 3102.5
## - n_stage                 2   2974.9 3107.7
## <none>                        2961.2 3110.6
## - race                    2   2980.1 3112.9
## - differentiate           3   2994.5 3119.0
## - estrogen_status         1   2978.0 3119.1
## - age                     1   2981.3 3122.4
## - progesterone_status     1   2982.3 3123.4
## - reginol_node_positive   1   2987.5 3128.6
## - sqrt_examined           1   2988.7 3129.8
## + marital_status          4   2953.4 3136.0
##
## Step:  AIC=3097.15
## status ~ race + n_stage + differentiate + a_stage + estrogen_status +
##     progesterone_status + ln_tumor + sqrt_examined + reginol_node_positive +
##     age
##
##                          Df Deviance    AIC
## - a_stage                 1   2974.2 3090.4
## - n_stage                 2   2987.1 3095.0
## <none>                        2972.7 3097.2
## - race                    2   2991.8 3099.7
## - estrogen_status         1   2989.9 3106.1
## - ln_tumor                1   2989.9 3106.1
## - differentiate           3   3007.4 3107.0
## - progesterone_status     1   2993.3 3109.5
## - age                     1   2993.4 3109.7
## + t_stage                 3   2961.2 3110.6
## - reginol_node_positive   1   2999.2 3115.4
## - sqrt_examined           1   2999.9 3116.1
## + marital_status          4   2964.7 3122.4
##
## Step:  AIC=3090.36
## status ~ race + n_stage + differentiate + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + reginol_node_positive + age
##
##                          Df Deviance    AIC
## - n_stage                 2   2989.3 3088.9
## <none>                        2974.2 3090.4
## - race                    2   2993.3 3092.9
## + a_stage                 1   2972.7 3097.2
## - estrogen_status         1   2991.9 3099.8
## - differentiate           3   3008.5 3099.8
## - ln_tumor                1   2992.1 3100.0
## + t_stage                 3   2961.2 3102.3
## - progesterone_status     1   2994.5 3102.4
## - age                     1   2994.7 3102.6
## - reginol_node_positive   1   3000.4 3108.3
## - sqrt_examined           1   3001.7 3109.6
## + marital_status          4   2966.0 3115.4
```

```
##
## Step:  AIC=3088.9
## status ~ race + differentiate + estrogen_status + progesterone_status +
##     ln_tumor + sqrt_examined + reginol_node_positive + age
##
##                         Df Deviance    AIC
## <none>                       2989.3 3088.9
## + n_stage                2   2974.2 3090.4
## - race                   2   3009.5 3092.5
## + a_stage                1   2987.1 3095.0
## - estrogen_status        1   3008.0 3099.3
## + t_stage                3   2975.1 3099.6
## - age                    1   3009.9 3101.2
## - progesterone_status    1   3010.3 3101.6
## - differentiate          3   3027.0 3101.7
## - ln_tumor               1   3013.8 3105.1
## - sqrt_examined          1   3015.5 3106.8
## + marital_status         4   2981.4 3114.2
## - reginol_node_positive  1   3146.2 3237.5
```

```
summary(bic_selected_bin_model)
```

```
##
## Call:
## glm(formula = status ~ race + differentiate + estrogen_status +
##     progesterone_status + ln_tumor + sqrt_examined + reginol_node_positive +
##     age, family = binomial, data = newbc2)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.451459   0.448634   5.464 4.65e-08 ***
## race2                   -0.595023   0.157272  -3.783 0.000155 ***
## race3                    0.440768   0.201932   2.183 0.029054 *
## differentiate2           0.417805   0.103868   4.022 5.76e-05 ***
## differentiate3           0.957010   0.190932   5.012 5.38e-07 ***
## differentiate4          -0.916981   0.506763  -1.809 0.070375 .
## estrogen_status1         0.763842   0.175498   4.352 1.35e-05 ***
## progesterone_status1     0.596058   0.126629   4.707 2.51e-06 ***
## ln_tumor                -0.366801   0.075028  -4.889 1.01e-06 ***
## sqrt_examined            0.248487   0.048812   5.091 3.57e-07 ***
## reginol_node_positive -0.109570   0.008846 -12.386  < 2e-16 ***
## age                     -0.024373   0.005417  -4.500 6.80e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3444.7  on 4023  degrees of freedom
## Residual deviance: 2989.3  on 4012  degrees of freedom
## AIC: 3013.3
##
## Number of Fisher Scoring iterations: 5
```

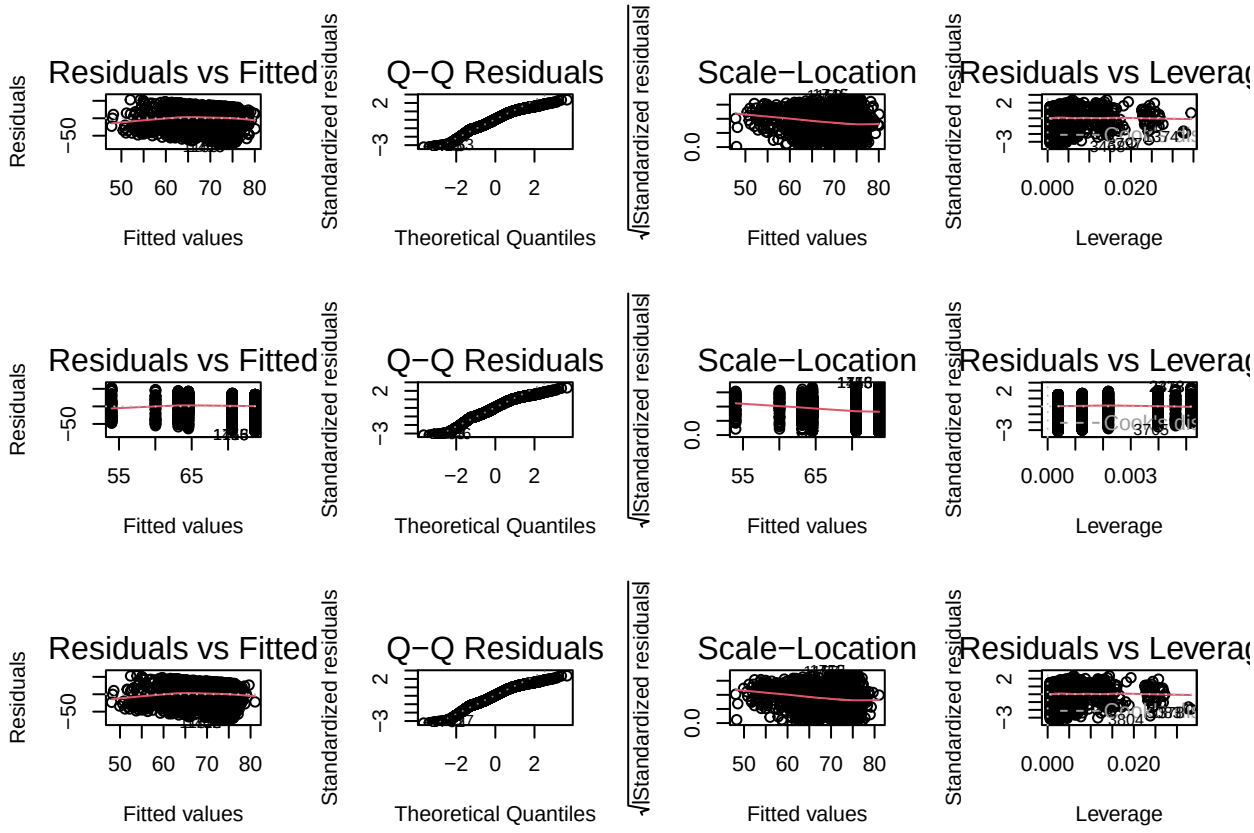
In the binary model, the final model according to the AIC selection is status ~ race + t_stage + n_stage +

differentiate + estrogen_status + progesterone_status + sqrt_examined + reginol_node_positive + age.

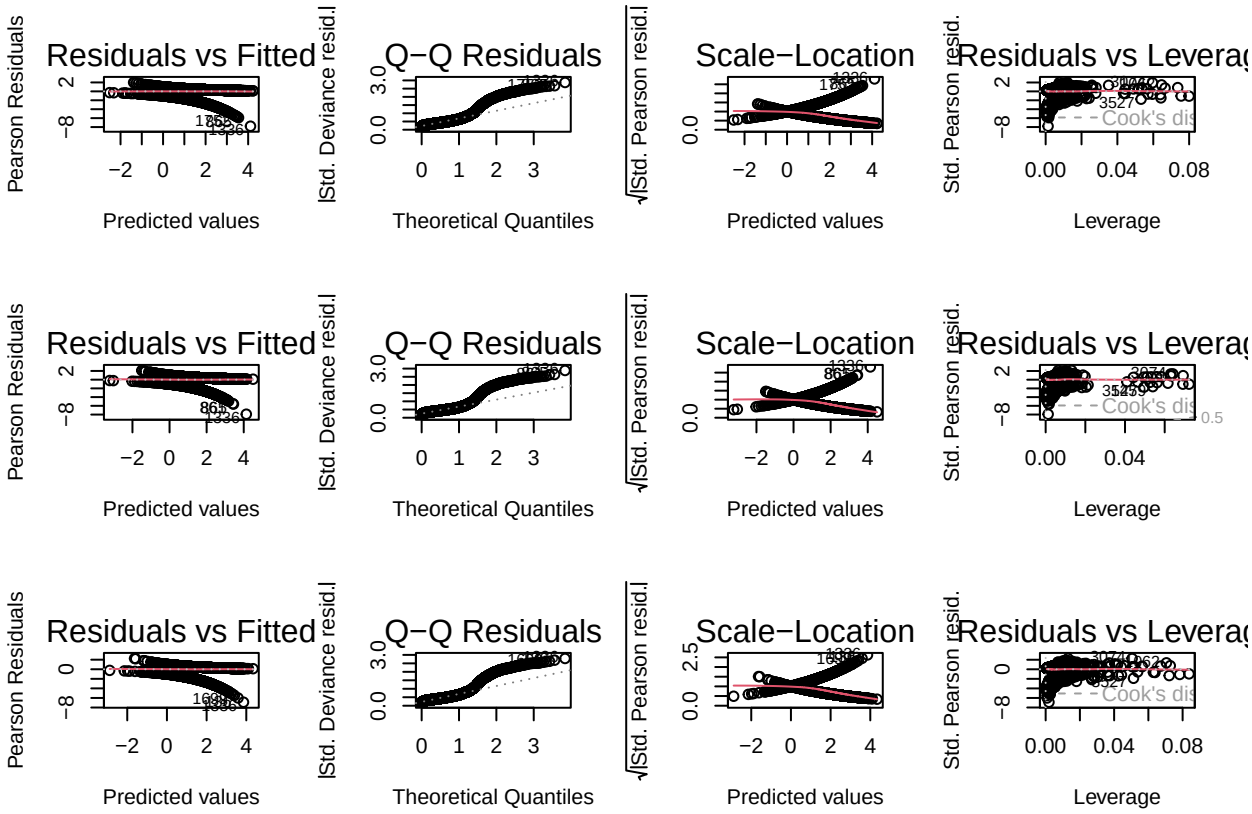
The BIC had the selection of model status ~ race + differentiate + estrogen_status + progesterone_status + ln_tumor + sqrt_examined + reginol_node_positive + age.

**Plots to check the model assumptions:**

```
par(mfrow = c(3,4))
plot(aic_selected_num_model)
plot(bic_selected_num_model)
plot(nummodel_partial_3)
```




```
par(mfrow = c(3,4))
plot(aic_selected_bin_model)
plot(bic_selected_bin_model)
plot(binmodel_partial_1)
```




**VIF**

3.0 transformation edited 3.1 interaction transformation ?? 3.2 partial test 3.3 diagnostic boxcox 4. Stepwise: forward/ backward /AIC 5. final model 6. model assumption (check multicolinearity (VIF)) 7. cross validation