P8130_final_project

Leonor Rui

2024-12-03

Appendix

- Data Import

```
survival_df = read_csv("data/Project_2_data.csv") |>
  janitor::clean_names()
```

- Data Description

```
str(survival_df)
```

```
## spc_tbl_ [4,024 x 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age                   : num [1:4024] 68 50 58 58 47 51 51 40 40 69 ...
##  $ race                  : chr [1:4024] "White" "White" "White" "White" ...
##  $ marital_status        : chr [1:4024] "Married" "Married" "Divorced" "Married" ...
##  $ t_stage               : chr [1:4024] "T1" "T2" "T3" "T1" ...
##  $ n_stage               : chr [1:4024] "N1" "N2" "N3" "N1" ...
##  $ x6th_stage            : chr [1:4024] "IIA" "IIIA" "IIIC" "IIA" ...
##  $ differentiate         : chr [1:4024] "Poorly differentiated" "Moderately differentiated" "Moderat
##  $ grade                 : chr [1:4024] "3" "2" "2" "3" ...
##  $ a_stage               : chr [1:4024] "Regional" "Regional" "Regional" "Regional" ...
##  $ tumor_size            : num [1:4024] 4 35 63 18 41 20 8 30 103 32 ...
##  $ estrogen_status       : chr [1:4024] "Positive" "Positive" "Positive" "Positive" ...
##  $ progesterone_status   : chr [1:4024] "Positive" "Positive" "Positive" "Positive" ...
##  $ regional_node_examined: num [1:4024] 24 14 14 2 3 18 11 9 20 21 ...
##  $ reginol_node_positive : num [1:4024] 1 5 7 1 1 2 1 1 18 12 ...
##  $ survival_months       : num [1:4024] 60 62 75 84 50 89 54 14 70 92 ...
##  $ status                : chr [1:4024] "Alive" "Alive" "Alive" "Alive" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Age = col_double(),
##   ..   Race = col_character(),
##   ..   ‘Marital Status‘ = col_character(),
##   ..   ‘T Stage‘ = col_character(),
##   ..   ‘N Stage‘ = col_character(),
##   ..   ‘6th Stage‘ = col_character(),
##   ..   differentiate = col_character(),
##   ..   Grade = col_character(),
##   ..   ‘A Stage‘ = col_character(),
##   ..   ‘Tumor Size‘ = col_double(),
##   ..   ‘Estrogen Status‘ = col_character(),
##   ..   ‘Progesterone Status‘ = col_character(),
##   ..   ‘Regional Node Examined‘ = col_double(),
##   ..   ‘Reginol Node Positive‘ = col_double(),
##   ..   ‘Survival Months‘ = col_double(),
##   ..   Status = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Numeric variables include age, tumor_size, regional_node_examined, reginol_node_positive, and survival_months.

These are continuous variables that can be used for our later regression analysis.

Categorical variables include race, marital_status, t_stage, n_stage, x6th_stage, differentiate, grade, a_stage, estrogen_status, progesterone_status, and status.

Then we will convert these variables into factors.

```
survival_df = survival_df |>
  mutate(
    race = factor(race),
    marital_status = factor(marital_status),
    t_stage = factor(t_stage),
    n_stage = factor(n_stage),
    x6th_stage = factor(x6th_stage),
    differentiate = factor(differentiate),
    grade = factor(grade),
    a_stage = factor(a_stage),
    estrogen_status = factor(estrogen_status),
    progesterone_status = factor(progesterone_status),
    status = factor(status)
  )
```

```
summary(survival_df)
```

```
##       age           race       marital_status t_stage   n_stage   x6th_stage
##  Min.   :30.00   Black: 291   Divorced : 486   T1:1603   N1:2732   IIA :1305
##  1st Qu.:47.00   Other: 320   Married  :2643   T2:1786   N2: 820   IIB :1130
##  Median :54.00   White:3413   Separated:  45   T3: 533   N3: 472   IIIA:1050
##  Mean   :53.97                Single   : 615   T4: 102             IIIB:  67
##  3rd Qu.:61.00                Widowed  : 235                       IIIC: 472
##  Max.   :69.00
##                    differentiate                   grade          a_stage
##  Moderately differentiated:2351   1                    : 543   Distant :  92
##  Poorly differentiated    :1111   2                    :2351   Regional:3932
##  Undifferentiated         :  19   3                    :1111
##  Well differentiated      : 543   anaplastic; Grade IV:  19
##
##
##    tumor_size     estrogen_status progesterone_status regional_node_examined
##  Min.   :  1.00   Negative: 269   Negative: 698       Min.   : 1.00
##  1st Qu.: 16.00   Positive:3755   Positive:3326       1st Qu.: 9.00
##  Median : 25.00                                       Median :14.00
##  Mean   : 30.47                                       Mean   :14.36  
##  3rd Qu.: 38.00                                       3rd Qu.:19.00  
##  Max.   :140.00                                       Max.   :61.00  
##  reginol_node_positive survival_months   status    
##  Min.   : 1.000        Min.   :  1.0   Alive:3408  
##  1st Qu.: 1.000        1st Qu.: 56.0   Dead : 616  
##  Median : 2.000        Median : 73.0               
##  Mean   : 4.158        Mean   : 71.3               
##  3rd Qu.: 5.000        3rd Qu.: 90.0               
##  Max.   :46.000        Max.   :107.0
```

The majority of patients in the dataset are White, accounting for approximately 84.82% of the total population. Black patients make up 7.23%, and patients classified as "Other" constitute 7.95%. This imbalance suggests that the dataset is heavily skewed towards White patients, which could influence the generalizability of the findings to other racial groups.

The wide range of values in variables such as `tumor_size`, `regional_node_examined`, and `survival_months` indicates the need to explore relationships and their potential nonlinearities with survival, giving us a possible analytical regression model.

```
colSums(is.na(survival_df))
```

```
##                    age                   race         marital_status 
##                      0                      0                      0 
##                t_stage                n_stage             x6th_stage 
##                      0                      0                      0 
##          differentiate                  grade                a_stage 
##                      0                      0                      0 
##             tumor_size        estrogen_status    progesterone_status 
##                      0                      0                      0 
## regional_node_examined  reginol_node_positive        survival_months 
##                      0                      0                      0 
##                 status 
##                      0
```

We can conclude that no missing values are present in this dataset across all variables.

```
survival_df |>
  group_by(differentiate, race) |>
  summarise(count = n(), .groups = "drop") |>
  pivot_wider(
    names_from = differentiate,
    values_from = count,
    values_fill = list(count = 0)
  )
```

```
## # A tibble: 3 x 5
##   race  ‘Moderately differentiated‘ ‘Poorly differentiated‘ Undifferentiated
##   <fct>                       <int>                   <int>            <int>
## 1 Black                         141                     115                3
## 2 Other                         180                      94                0
## 3 White                        2030                     902               16
## # i 1 more variable: ‘Well differentiated‘ <int>
```

This table shows the frequency of different levels of `differentiate` by races.

```
survival_df |>
  group_by(x6th_stage, status) |>
  summarise(count = n(), .groups = "drop") |>
  pivot_wider(
    names_from = status,
    values_from = count
  )
```

```
## # A tibble: 5 x 3
##   x6th_stage Alive  Dead
##   <fct>      <int> <int>
## 1 IIA         1209    96
## 2 IIB          995   135
## 3 IIIA         866   184
## 4 IIIB          47    20
## 5 IIIC         291   181
```

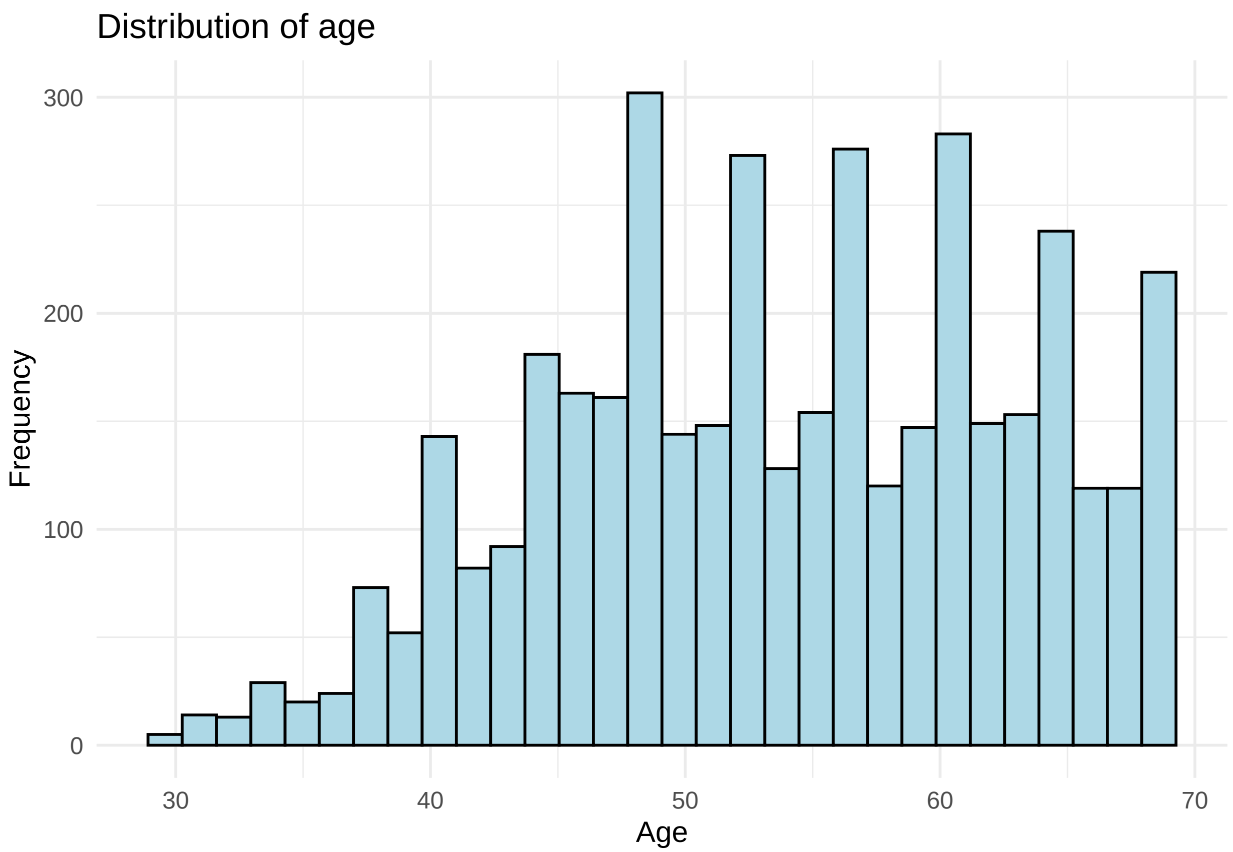
This table shows the frequency of different levels of `status` by 6th stage.

- Data Visualization

Distributions of the numeric variables

Distribution of age

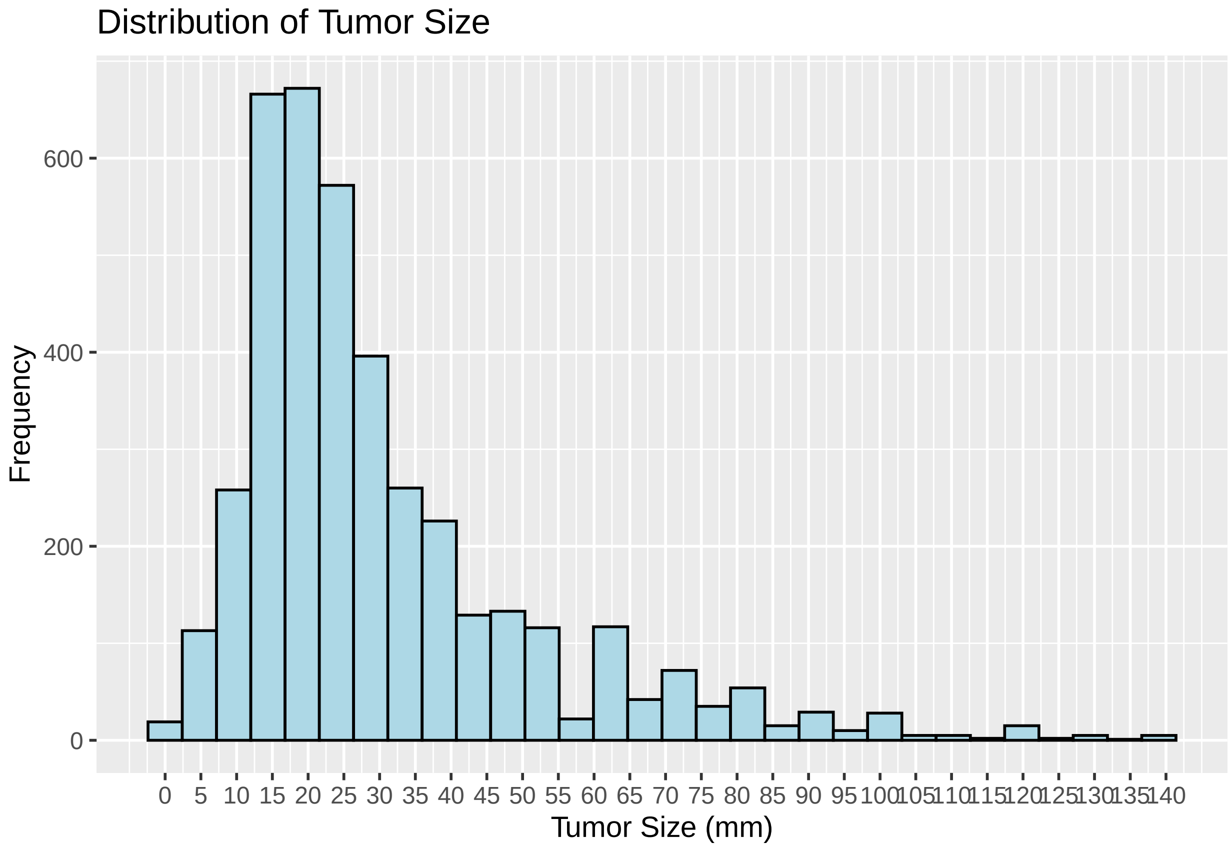
```
survival_df |>
  ggplot(aes(age)) +
  geom_histogram(fill = "light blue", color = "black") +
  theme_minimal() +
  labs(
    title = "Distribution of age",
    x = "Age",
    y = "Frequency"
  )
```



The histogram shows the age distribution of patients. Most patients are aged between 40 and 70 years. The data is well spread across middle and older age groups, making it possible for age-related analysis. Therefore, age will likely be a significant predictor for later analysis.

Distribution of tumor size

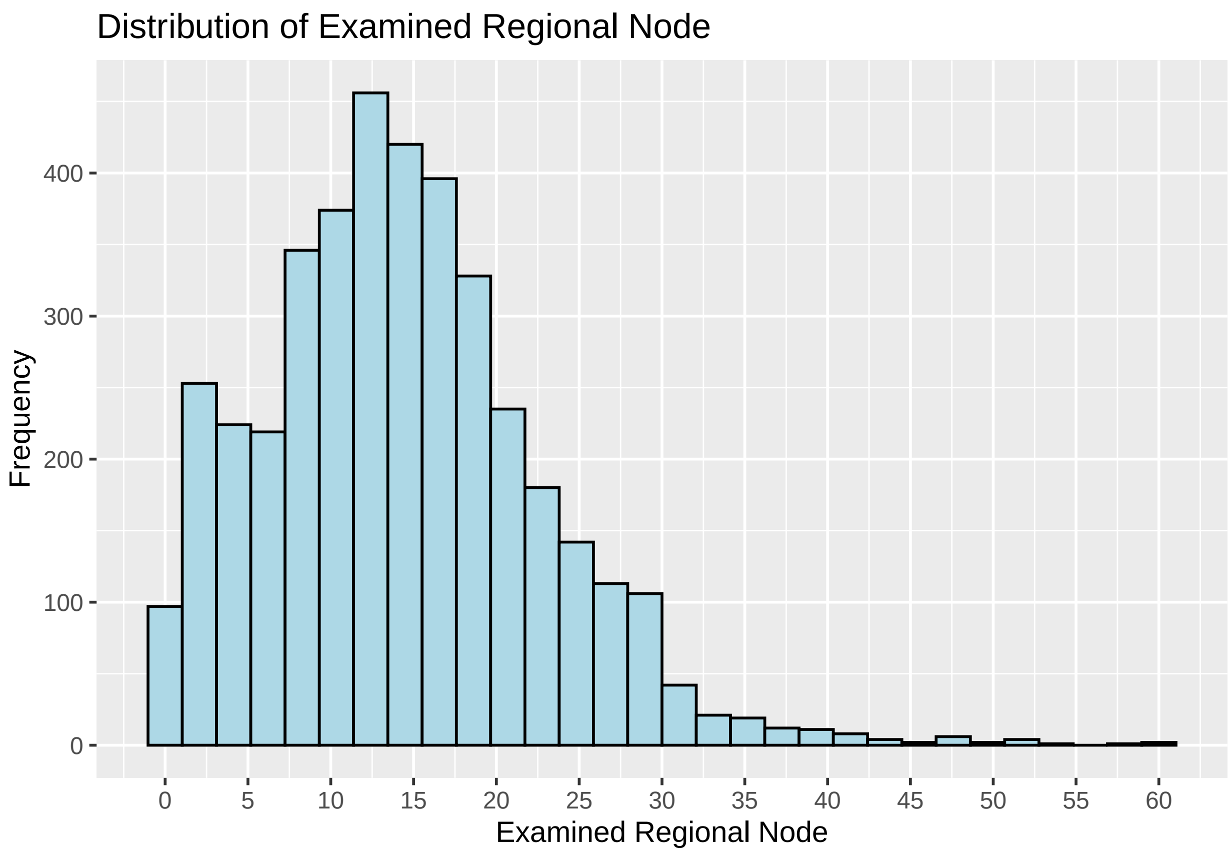
```r
ggplot(survival_df, aes(x = tumor_size)) +
  geom_histogram(fill = "light blue", color = "black") +
  scale_x_continuous(breaks = seq(0, max(survival_df$tumor_size, na.rm = TRUE), by = 5)) +
  labs(
    title = "Distribution of Tumor Size",
    x = "Tumor Size (mm)",
    y = "Frequency"
  )
```



This is the distribution of all tumor sizes, and most of the tumor sizes are smaller than 50 mm. We can find that the most frequent size is around 19 mm, followed by around 14 mm. This distribution is right-skewed, so we will use the log transformation for this variable.

Distribution of examined regional node

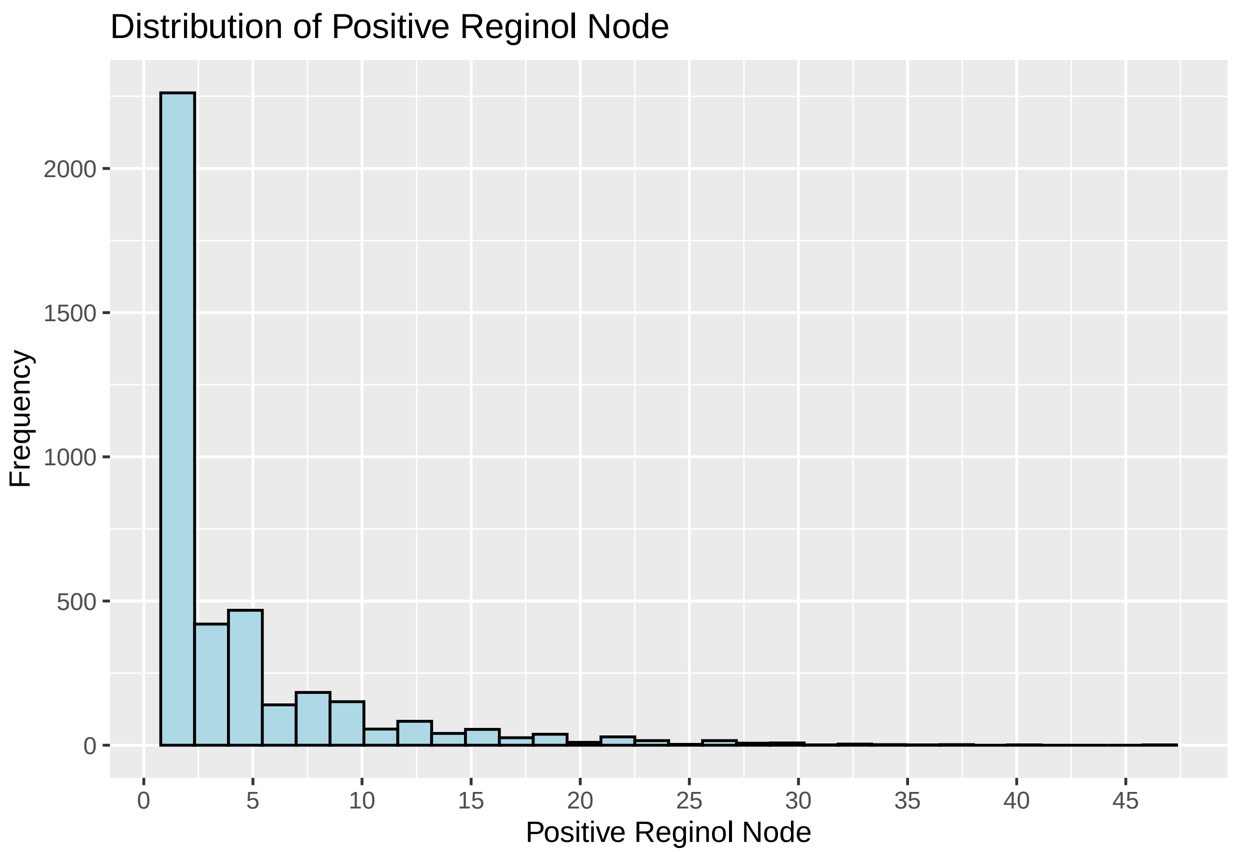
```
ggplot(survival_df, aes(x = regional_node_examined)) +
  geom_histogram(fill = "light blue", color = "black") +
  scale_x_continuous(breaks = seq(0, max(survival_df$regional_node_examined, na.rm = TRUE), by = 5)) +
  labs(
    title = "Distribution of Examined Regional Node",
    x = "Examined Regional Node",
    y = "Frequency"
  )
```



This plot maps the frequency of different number of examined regional nodes for each subject. The number of examined regional nodes for most subjects are smaller than 30, and the subjects with nearly 12 examined regional nodes are the most.

Distribution of positive regional node

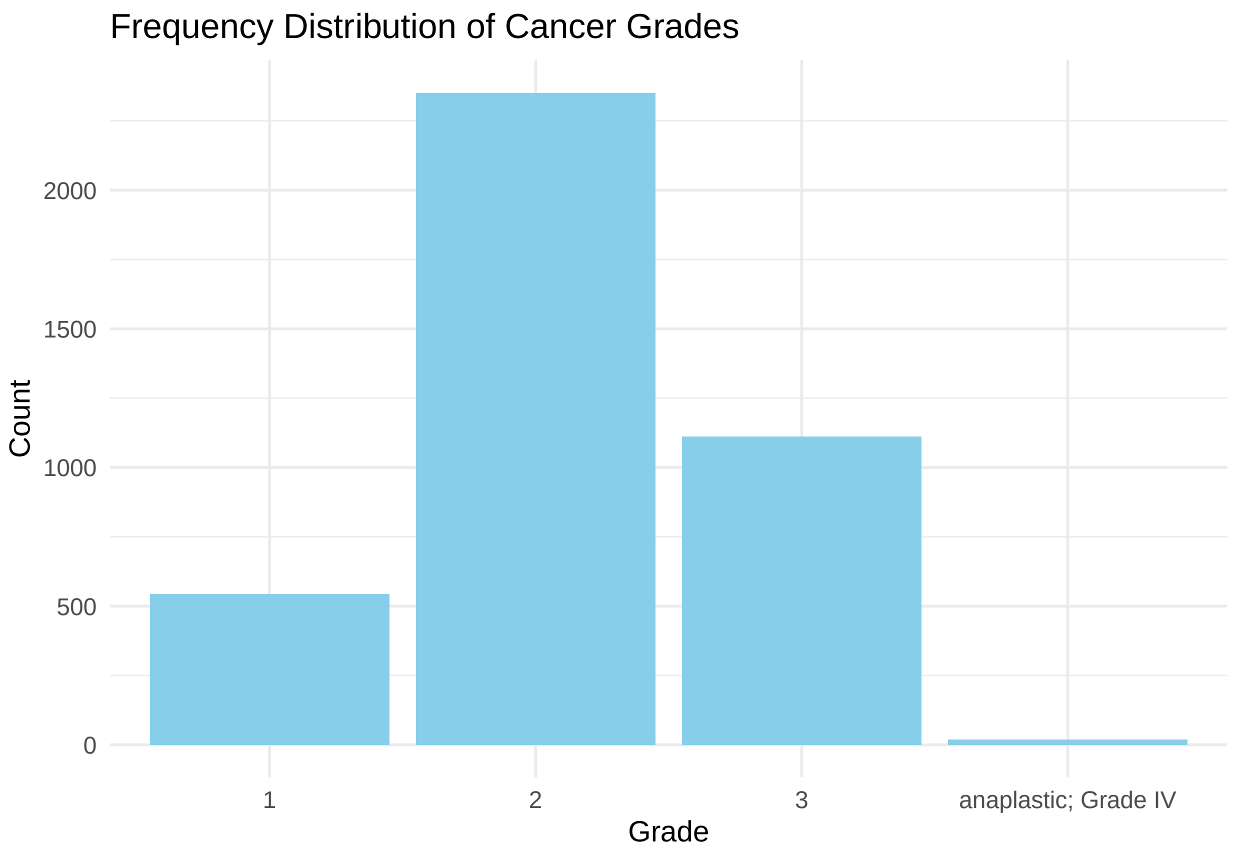
```
ggplot(survival_df, aes(x = reginol_node_positive)) +
  geom_histogram(fill = "light blue", color = "black") +
  scale_x_continuous(breaks = seq(0, max(survival_df$reginol_node_positive, na.rm = TRUE), by = 5)) +
  labs(
    title = "Distribution of Positive Reginol Node",
    x = "Positive Reginol Node",
    y = "Frequency"
  )
```



Then is the distribution of different number of positive reginol node for each subject. Over 2500 subjects only have 1 or 2 positive reginol nodes, which is the most frequent number of positive reginol nodes. It is strongly right-skewed, so we will use the log transformation for this variable.

Distribution of Cancer Grades

```
ggplot(survival_df, aes(x = grade)) +
    geom_bar(fill = "skyblue") +
    labs(title = "Frequency Distribution of Cancer Grades",
         x = "Grade",
         y = "Count") +
    theme_minimal()
```

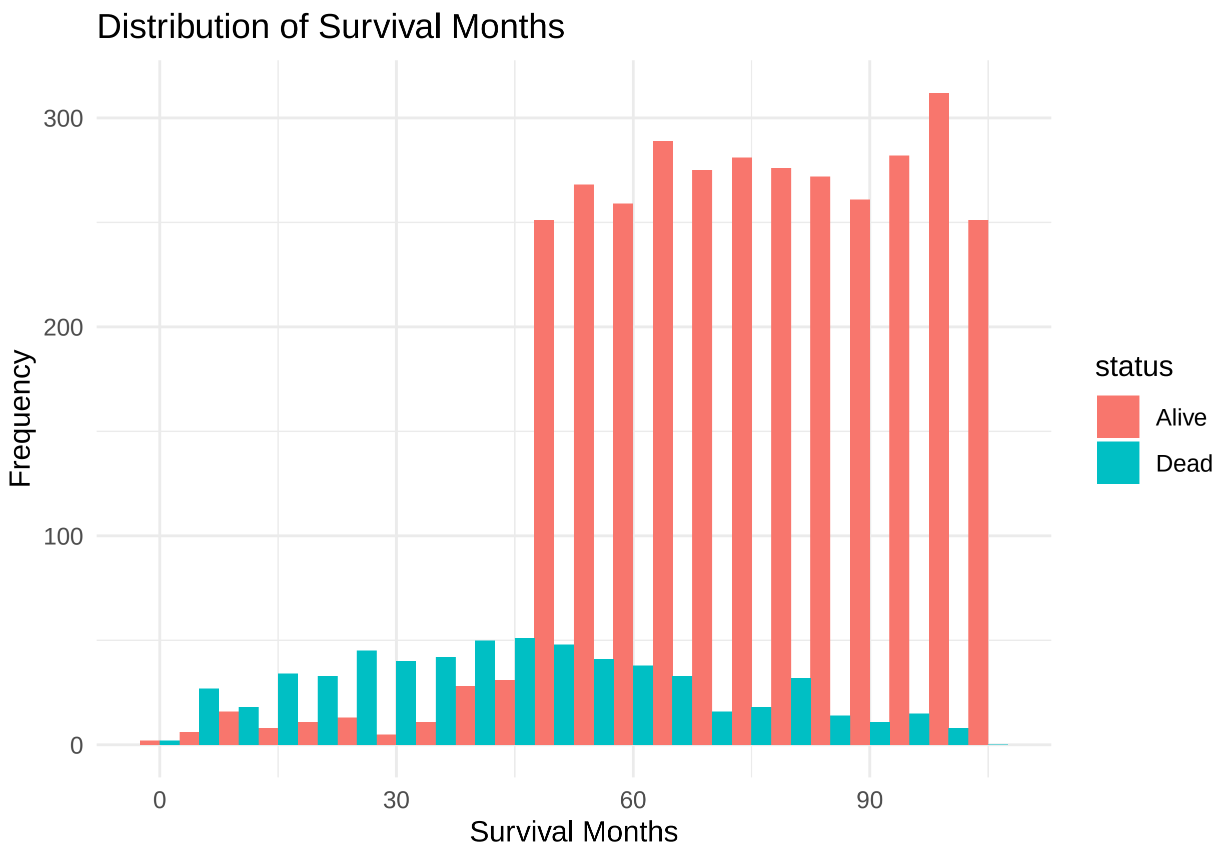


This bar chart provides an overview of how cancer cases are distributed across grades. Grade 2 represents the majority of cases, suggesting it is the most frequently observed grade, while Grade IV is exceedingly rare.

Bewteen Variables

Distribution of survival months by status

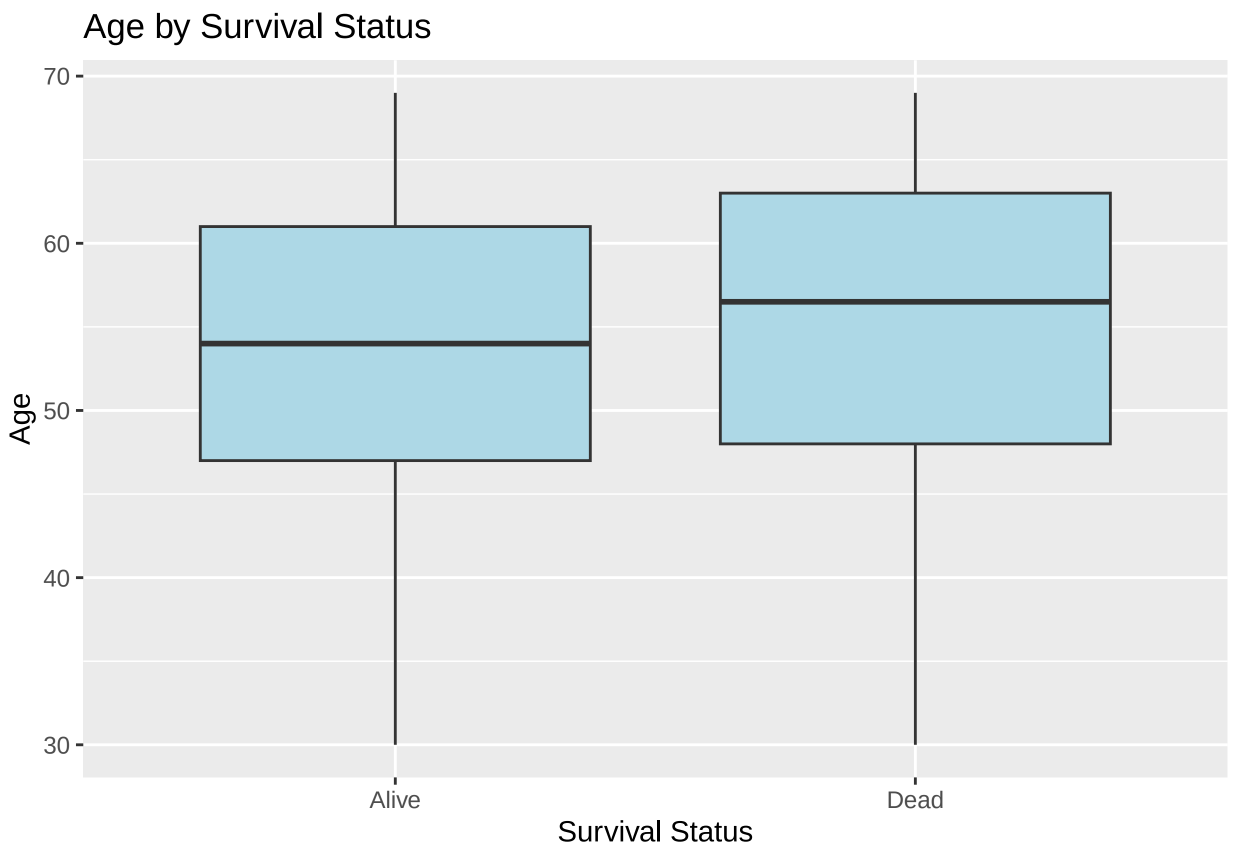
```
ggplot(survival_df, aes(x = survival_months, fill = status)) +
  geom_histogram(binwidth = 5, position = "dodge") +
  labs(title = "Distribution of Survival Months", x = "Survival Months", y = "Frequency") +
  theme_minimal()
```



The Dead group is concentrated in the shorter survival months, while the Alive group is predominant in longer survival months, particularly beyond 60 months.

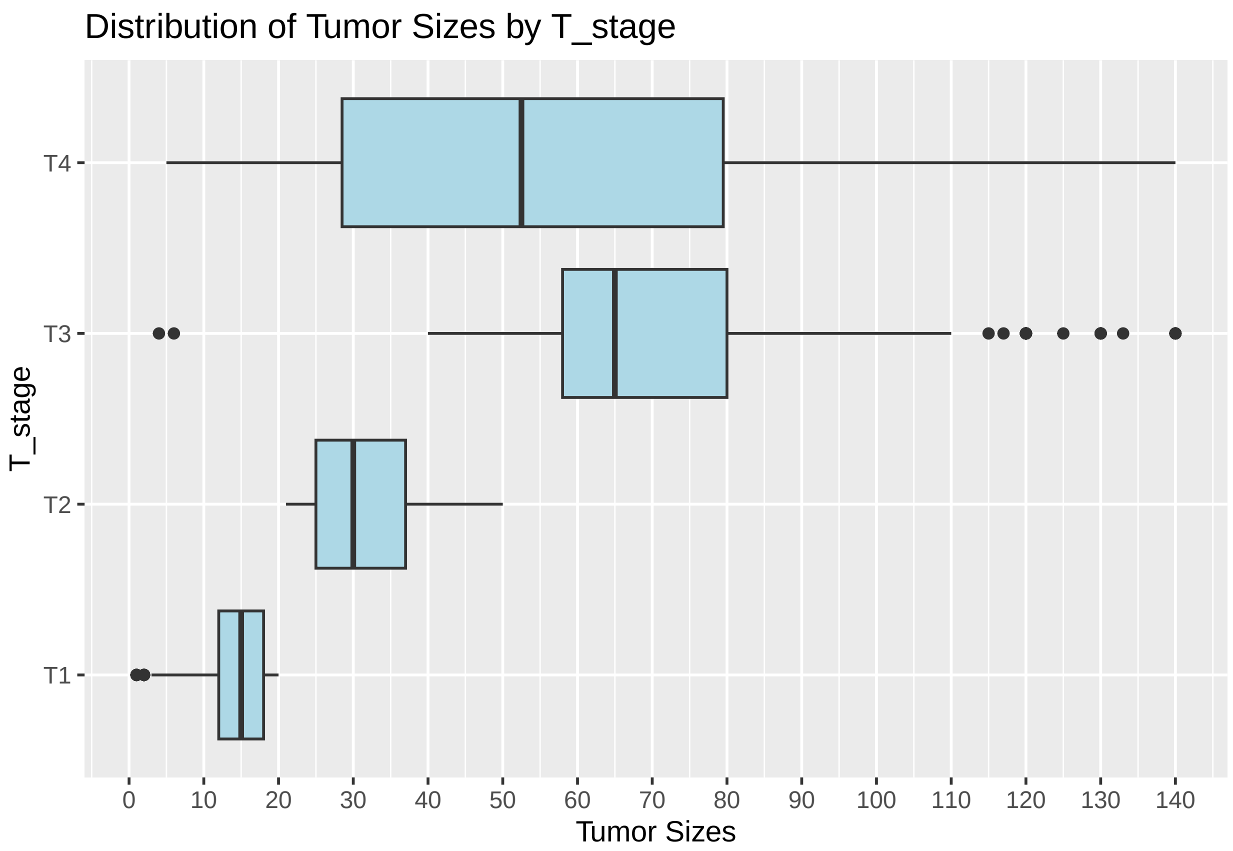
Distribution of age by survival status

```
ggplot(survival_df, aes(x = status, y = age)) +
  geom_boxplot(fill = "light blue") +
  labs(title = "Age by Survival Status", x = "Survival Status", y = "Age")
```



The distribution of the tumor sizes by t_stage

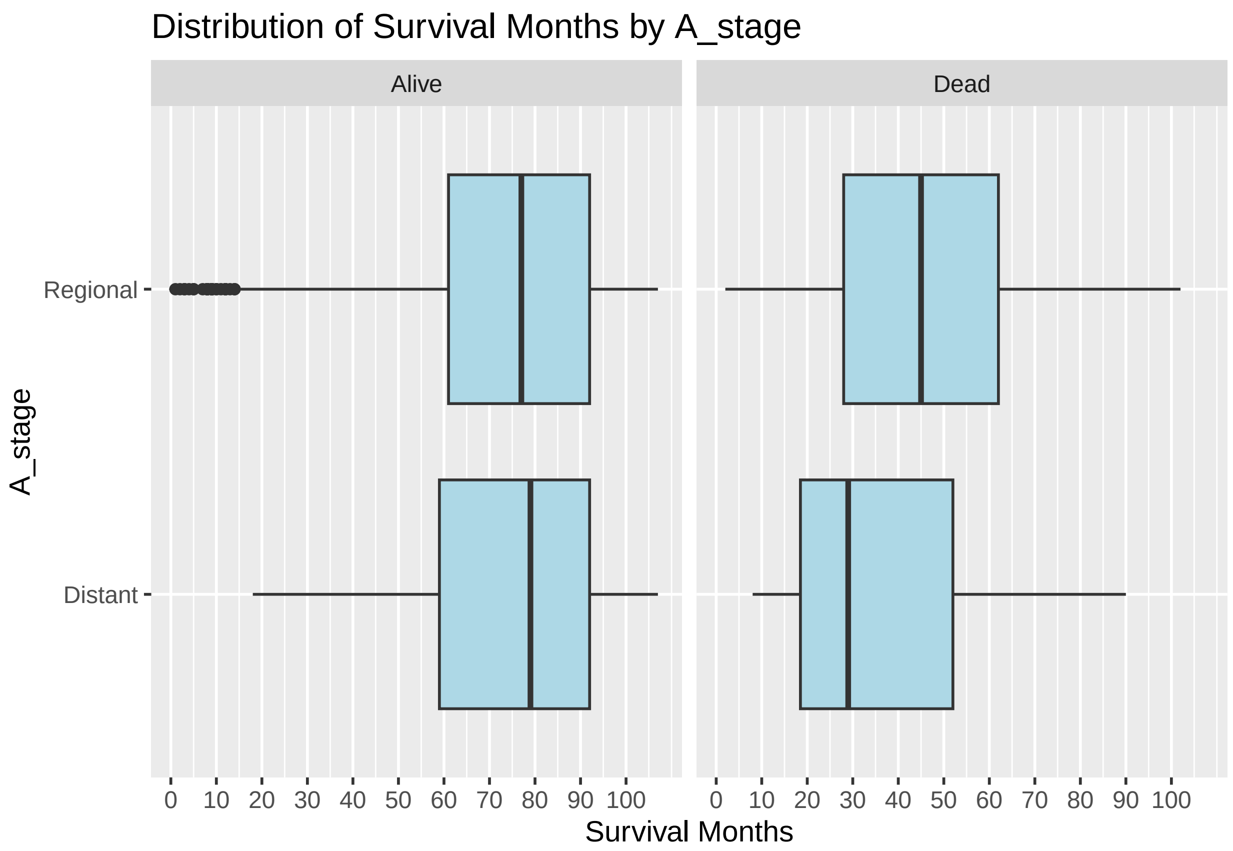
```r
ggplot(survival_df, aes(x = tumor_size, y = t_stage)) +
  geom_boxplot(fill = "light blue") +
  scale_x_continuous(breaks = seq(0, max(survival_df$tumor_size, na.rm = TRUE), by = 10)) +
  labs(
    title = "Distribution of Tumor Sizes by T_stage",
    x = "Tumor Sizes",
    y = "T_stage"
  )
```



In this plot, we explore the tumor size distribution at different T stages. From T1 to T3, as the stage changes, both the mean tumor sizes and IQR become larger. At T4 stage, the IQR of tumor sizes is much larger than others, and the mean size is smaller than the mean size at T3 stage. There are some outliers both ar T1 stage and T3 stage.

The distribution of survival months by a_stage based on status(alive/dead)

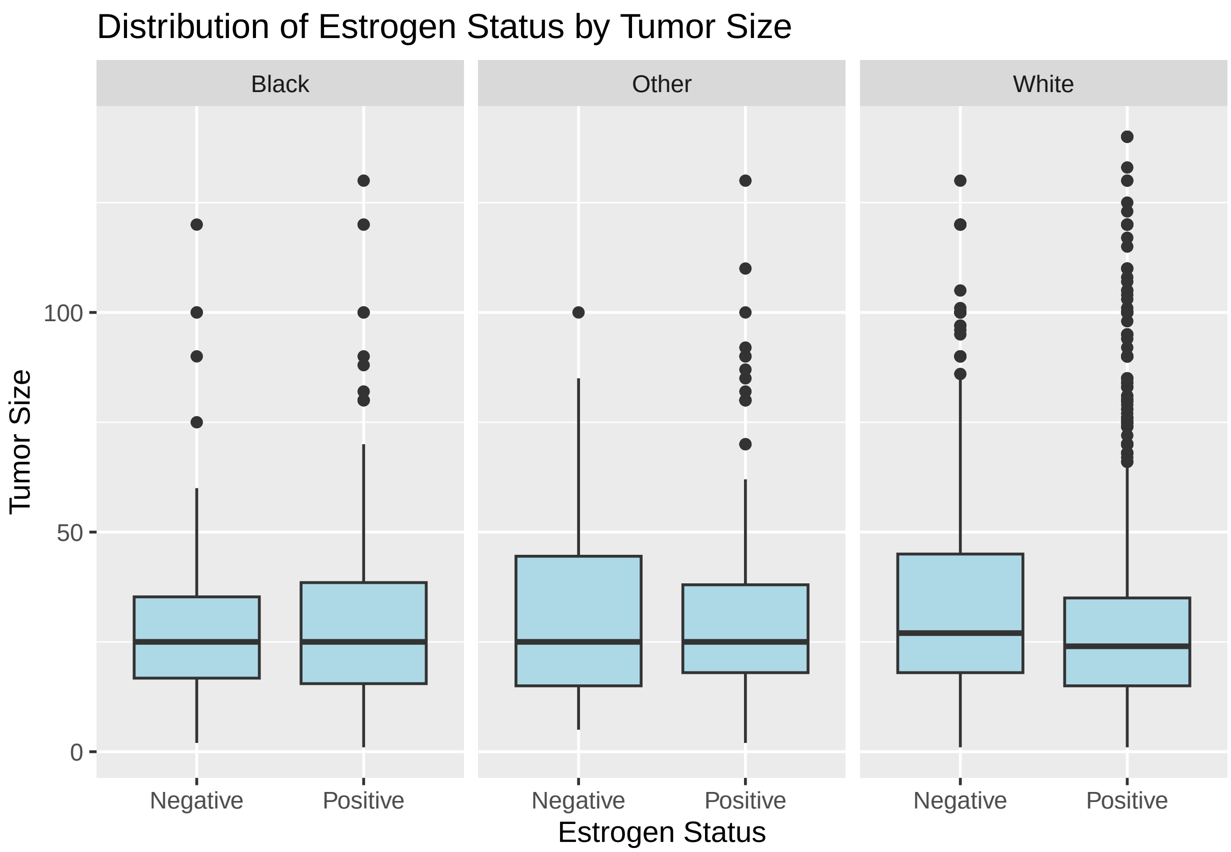
```r
ggplot(survival_df, aes(x = survival_months, y = a_stage)) +
  geom_boxplot(fill = "light blue") +
  scale_x_continuous(breaks = seq(0, max(survival_df$survival_months, na.rm = TRUE), by = 10)) +
  labs(
    title = "Distribution of Survival Months by A_stage",
    x = "Survival Months",
    y = "A_stage"
  ) +
  facet_grid(~ status)
```



Through this plot, we can find that subjects with Distant stage have fewer survival months than subjects with Regional stage. However, the IQR of the survival months of subjects with Distant stage is much larger than subjects with Regional stage.

Distribution of Estrogen Status by Tumor Size Based on race

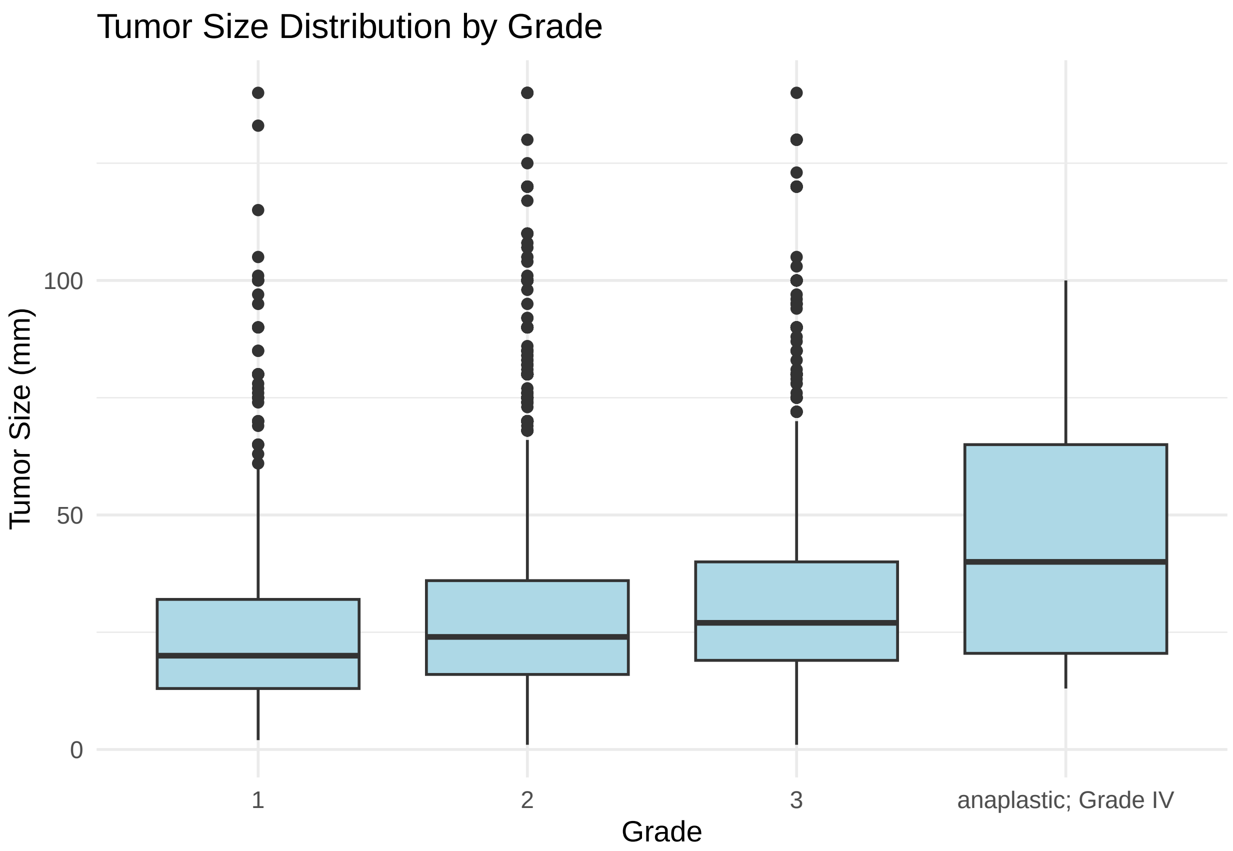
```
ggplot(survival_df, aes(x = progesterone_status, y = tumor_size)) +
  geom_boxplot(fill = "light blue") +
  labs(
    title = "Distribution of Estrogen Status by Tumor Size",
    x = "Estrogen Status",
    y = "Tumor Size"
  ) +
  facet_grid(~ race)
```



While the overall patterns are consistent, with Negative estrogen status generally associated with slightly larger tumor sizes, the variability and prevalence of outliers differ between groups. The White group shows the greatest spread in tumor size, while the Other group displays the least variability.

Tumor Size Distribution by Grade

```
ggplot(survival_df, aes(x = grade, y = tumor_size)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Tumor Size Distribution by Grade",
       x = "Grade",
       y = "Tumor Size (mm)") +
  theme_minimal()
```

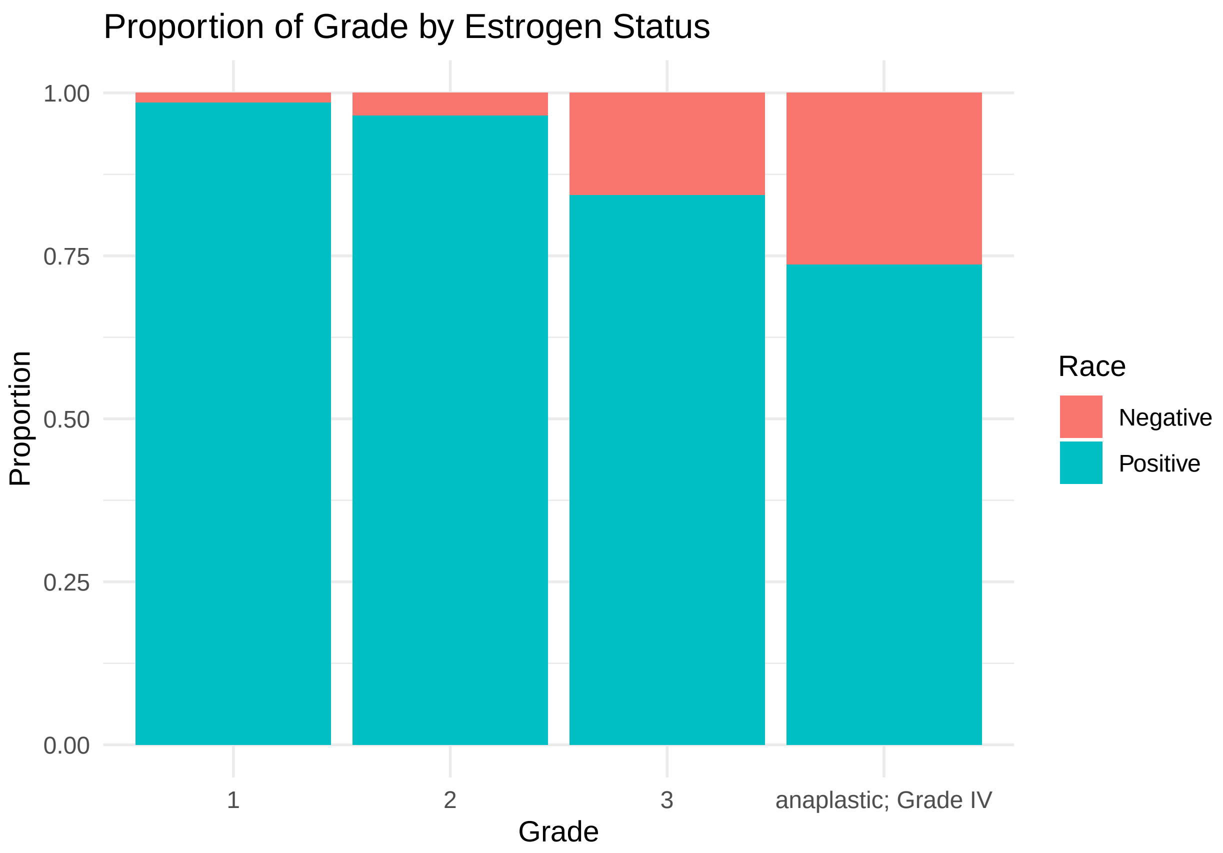


Lower grades (1–3) exhibit comparable tumor size distributions, with slight increases in variability as the grade increases.

Grade IV stands out due to its higher median and broader range, suggesting that more aggressive tumor grades are associated with larger tumor sizes.

Proportion of Grade by Estrogen Status

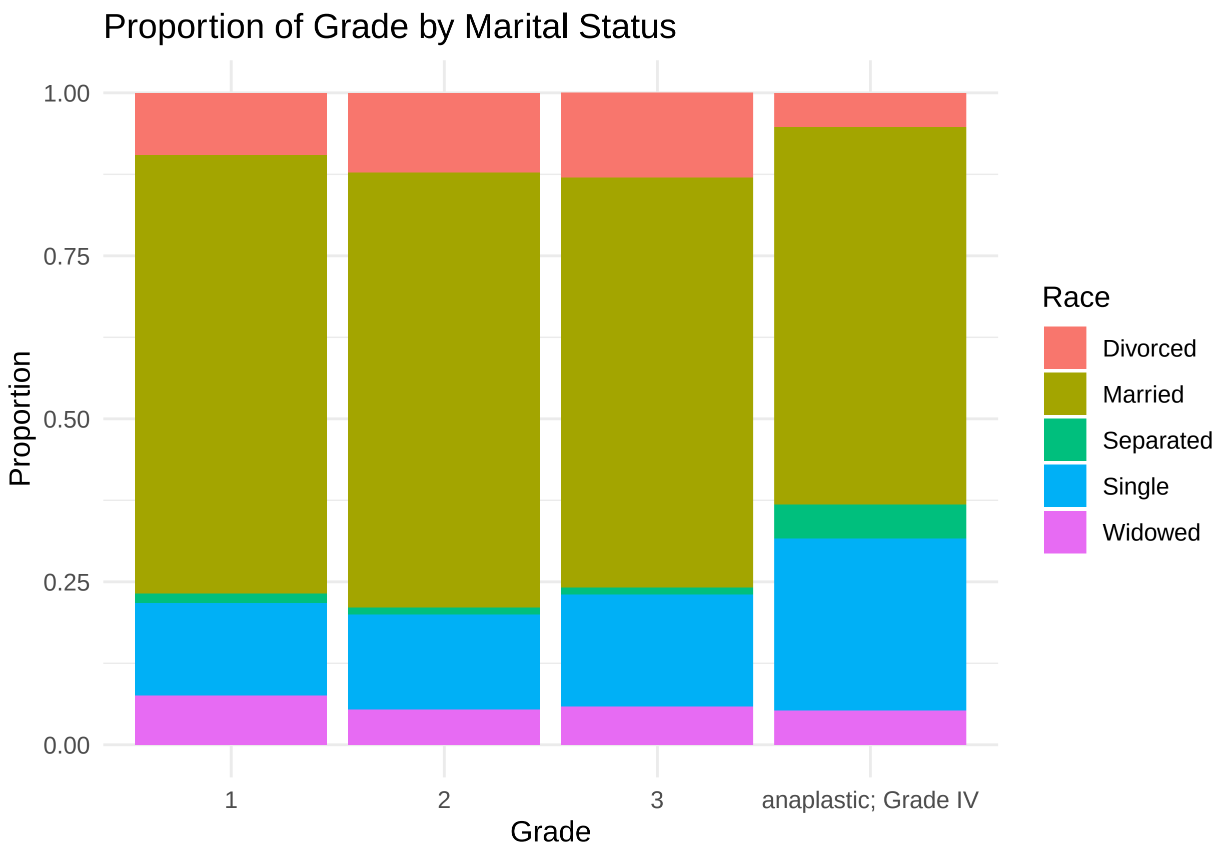
```
ggplot(survival_df, aes(x = grade, fill = estrogen_status)) +
    geom_bar(position = "fill") +
    labs(title = "Proportion of Grade by Estrogen Status",
         x = "Grade",
         y = "Proportion",
         fill = "Race") +
    theme_minimal()
```



As tumor grade increases, the proportion of Negative estrogen status gradually increases, becoming more prominent in the anaplastic Grade IV category. Conversely, the dominance of the Positive estrogen status decreases with higher tumor grades.

Proportion of Grade by Marital Status

```r
ggplot(survival_df, aes(x = grade, fill = marital_status)) +
    geom_bar(position = "fill") +
    labs(title = "Proportion of Grade by Marital Status",
         x = "Grade",
         y = "Proportion",
         fill = "Race") +
    theme_minimal()
```



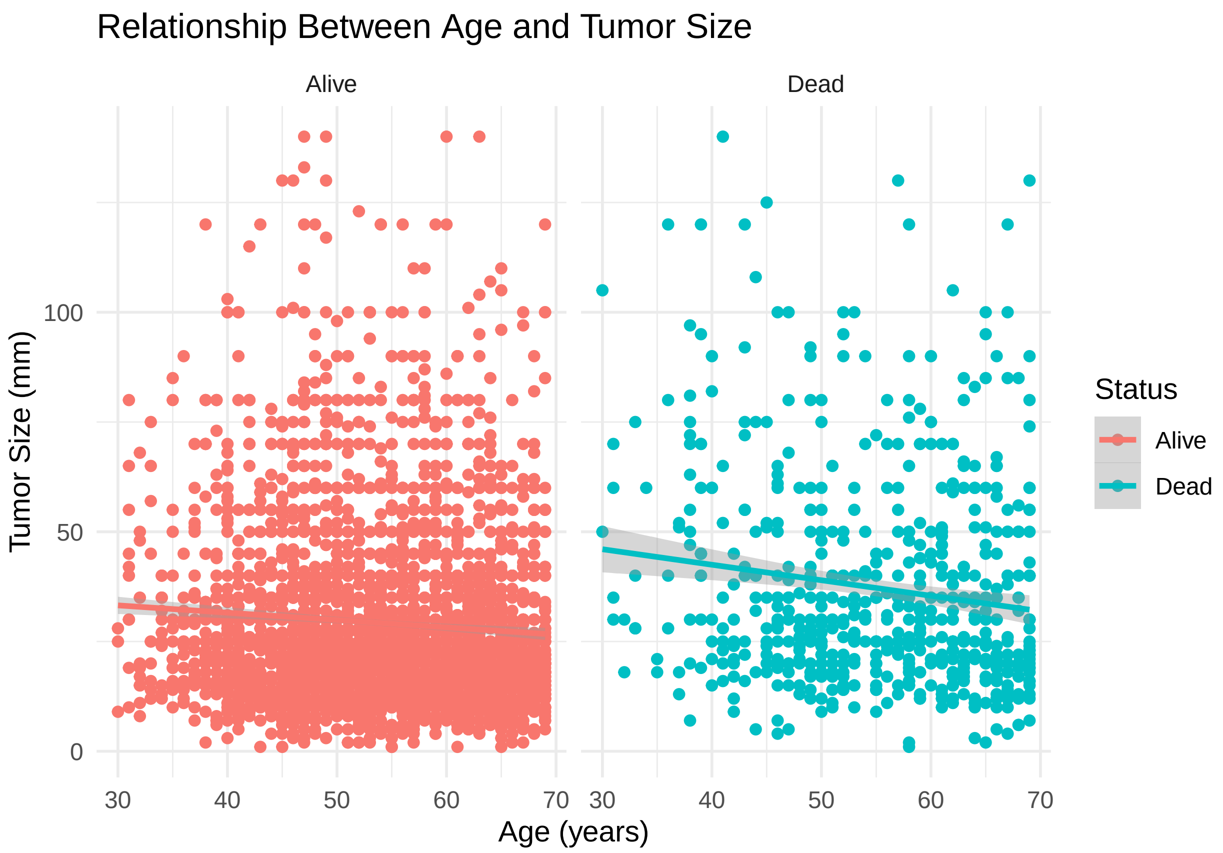
Across all grades, the “Married” group consistently constitutes the largest proportion of individuals, dominating every tumor grade category.

The “Single” group is the second-largest proportion in most grades, particularly Grades 2 and 3.

The “Widowed” group and “Divorced” group make up smaller proportions across all tumor grades.

Relationship Between Age and Tumor Size across status

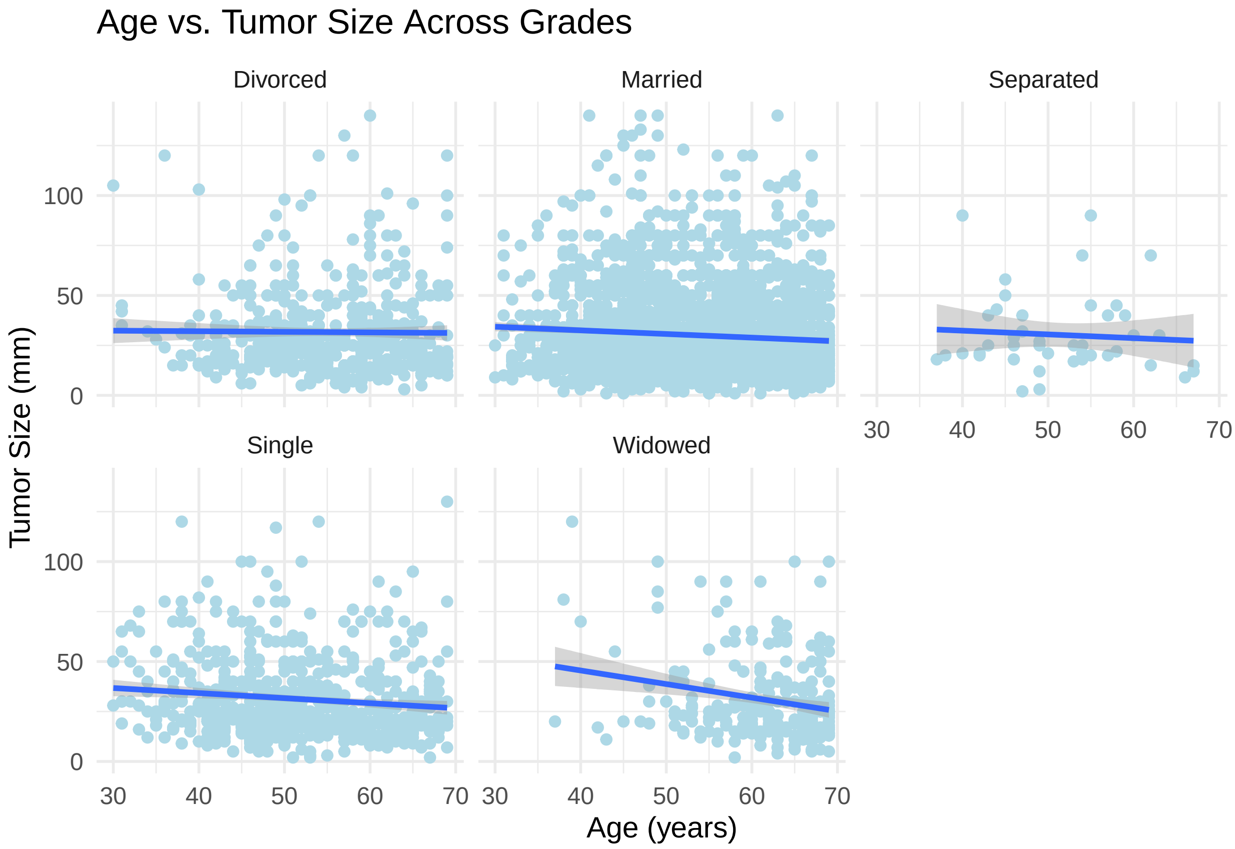
```
ggplot(survival_df, aes(x = age, y = tumor_size, color = status)) +
    geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ status) +
    labs(title = "Relationship Between Age and Tumor Size",
         x = "Age (years)",
         y = "Tumor Size (mm)",
         color = "Status") +
    theme_minimal()
```



This figure highlights the differences in tumor size distribution and trends with age between individuals who are alive and those who are deceased. While the "Alive" group shows no significant relationship between age and tumor size, the "Dead" group exhibits a pattern where larger tumors are associated with younger ages.

Age vs. Tumor Size Across Grades

```
ggplot(survival_df, aes(x = age, y = tumor_size)) +
    geom_point(color = "lightblue") +
  geom_smooth(method = "lm") +
    facet_wrap(~ marital_status) +
    labs(title = "Age vs. Tumor Size Across Grades",
         x = "Age (years)",
         y = "Tumor Size (mm)") +
    theme_minimal()
```



Divorced: Tumor size seems to remain fairly constant with age, as the trend line is relatively flat.

Married: A slight negative trend is observable, suggesting that tumor size may decrease marginally with age.

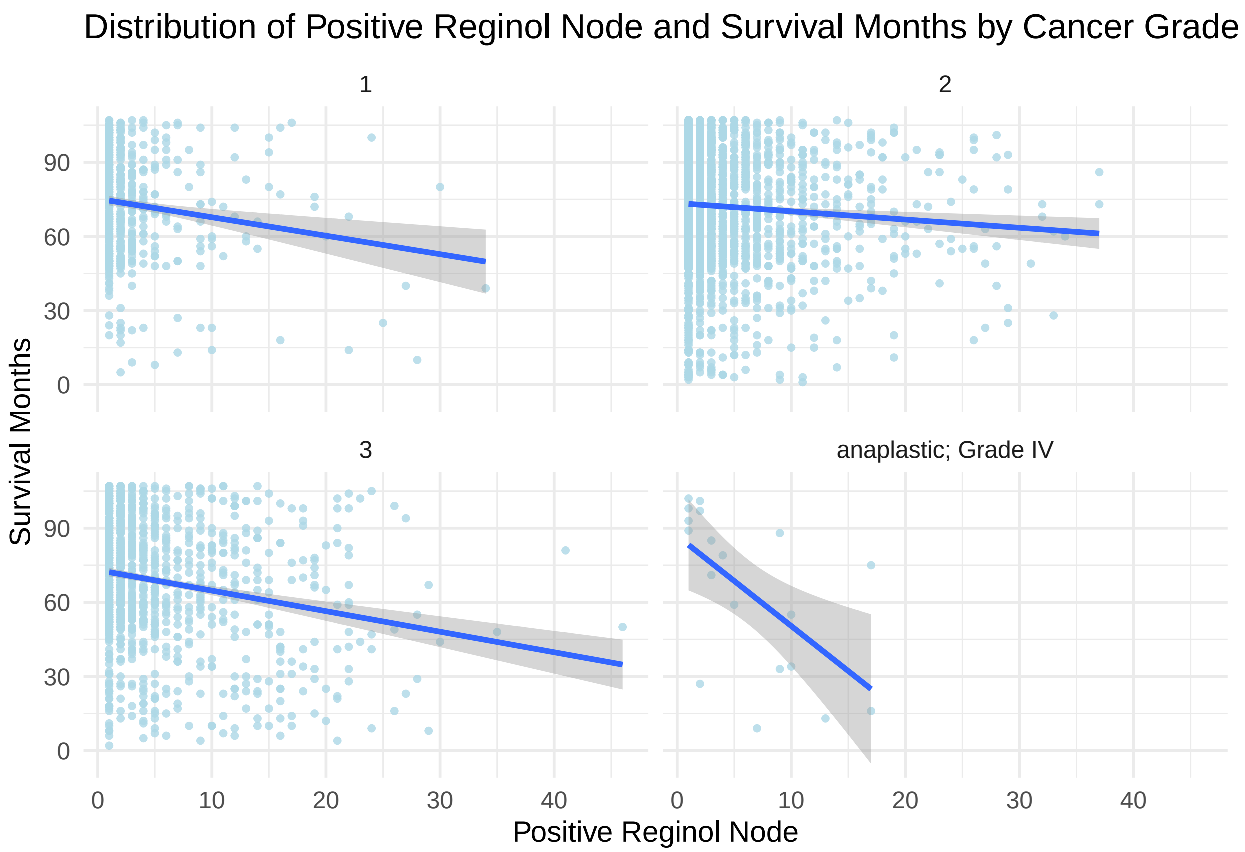
Separated: The data is sparse, but the trend shows a slightly negative relationship, with wide confidence intervals due to fewer observations.

Single: A modest negative trend is observed, indicating a potential decline in tumor size with increasing age.

Widowed: A more apparent negative trend is evident compared to other groups, suggesting a stronger decrease in tumor size with age.

Positive Reginol Node vs Survival Months Across Cancer Grade

```r
ggplot(survival_df, aes(x = reginol_node_positive, y = survival_months)) +
  geom_point(color = "light blue", size = 0.8, alpha = 0.8)  +
  facet_wrap(.~grade) +
  geom_smooth(method = "lm") +
  labs(
    title = "Distribution of Positive Reginol Node and Survival Months by Cancer Grade",
    x = "Positive Reginol Node",
    y = "Survival Months"
  ) +
  theme_minimal()
```



According to the trend lines, as the cancer grade increases, the negative correlation between the number of positive reginol nodes and the survival months becomes stronger. At the Grade IV, the correlation is strong. AS the number of positive reginol nodes increases, the survival months will decrease.

Transformations

```r
survival_df = survival_df |>
  mutate(
    log_tumor_size = log(tumor_size),
    log_reginol_node_positive = log(reginol_node_positive)
  )
```

Since variables `tumor_size` and `reginol_node_positive` are skewed to the right, we need to use the log transformation and add new variables `log_tumor_size` and `log_reginol_node_positive` for further analysis.