P8130_final_project

Leonor Rui

2024-12-03

Appendix

- Data Import

```
survival_df = read_csv("data/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.
```

- 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" "Moderate
##  $ 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 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))
```

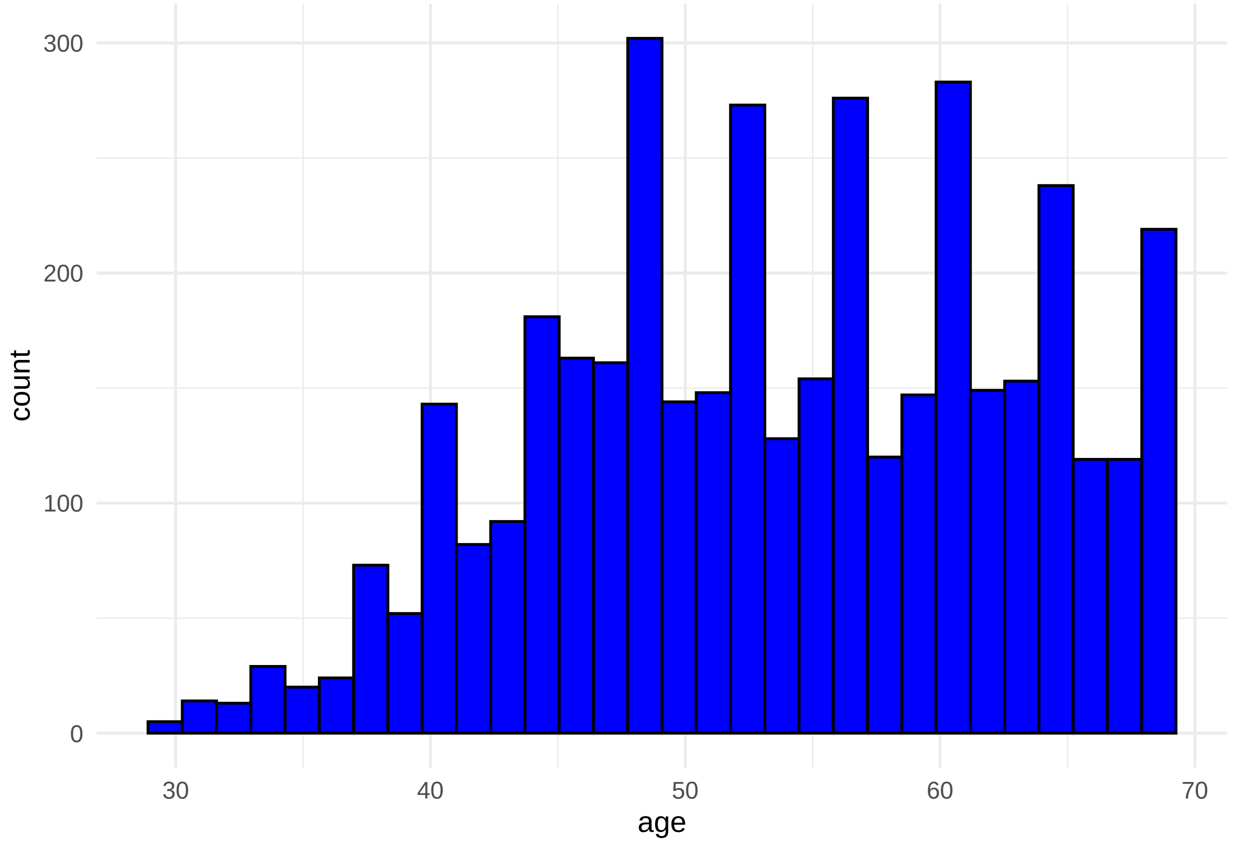
```
##                    age                   race         marital_status
##                      0                      0                      0
##                t_stage                n_stage             x6th_stage
##                      0                      0                      0
##          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.

- Data Visualization

```
survival_df |>
  ggplot(aes(age)) +
  geom_histogram(fill = "blue", color = "black") +
  theme_minimal()
```

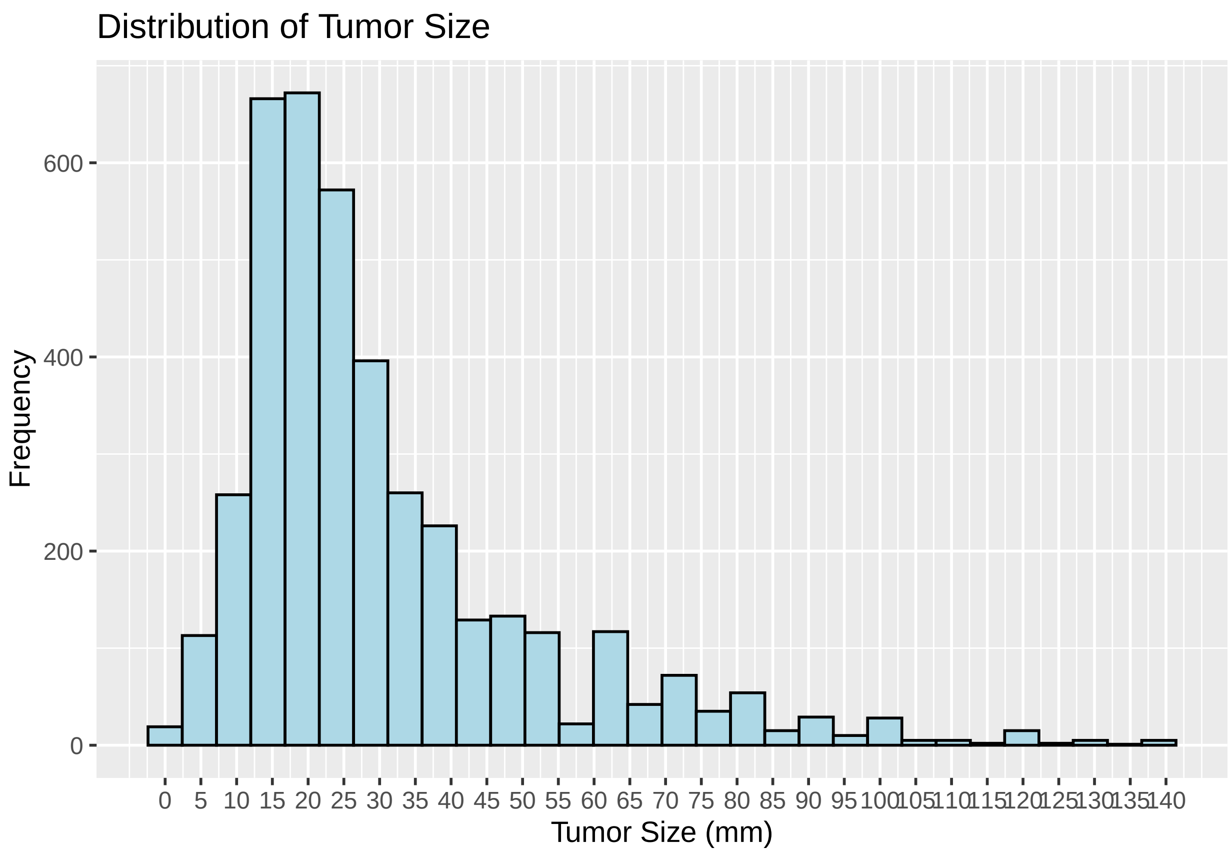
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.

```
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"
  )
```

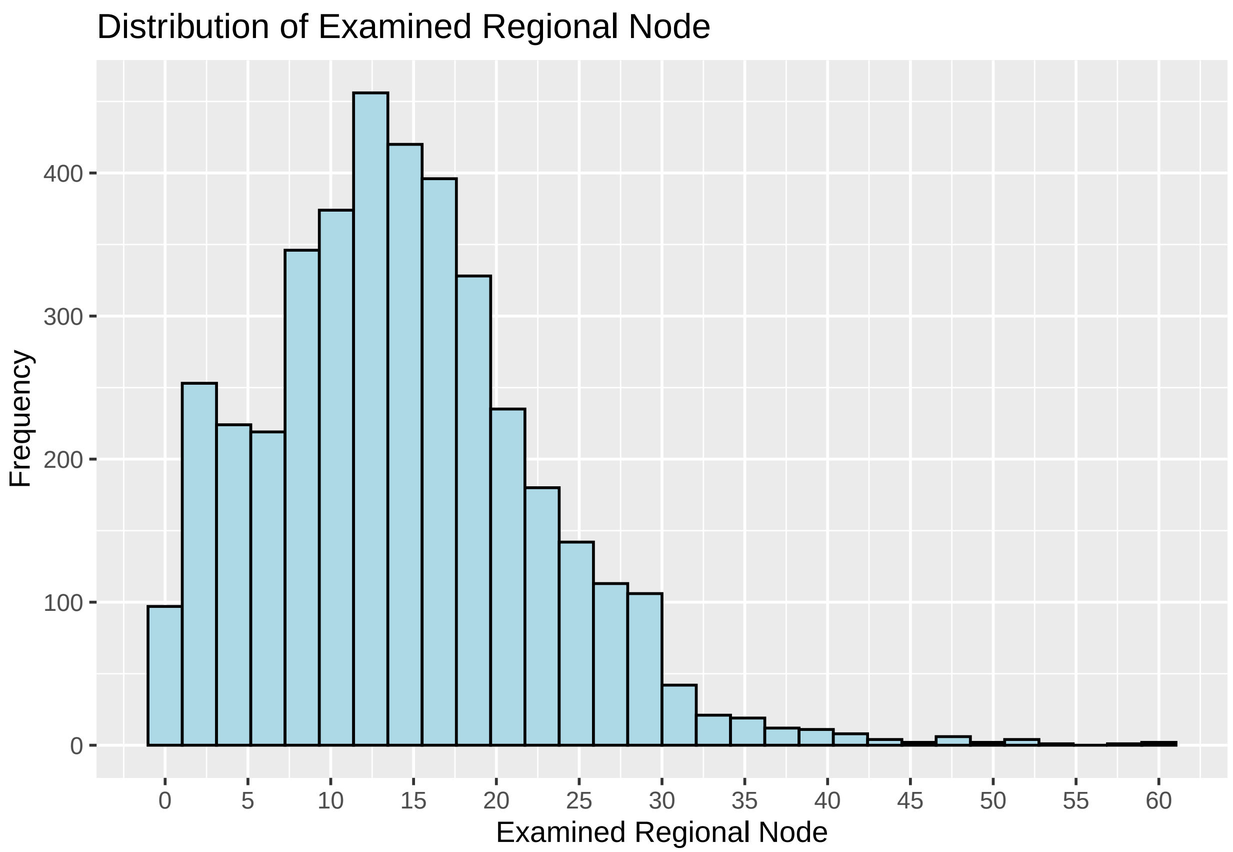
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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.

```
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"
  )
```

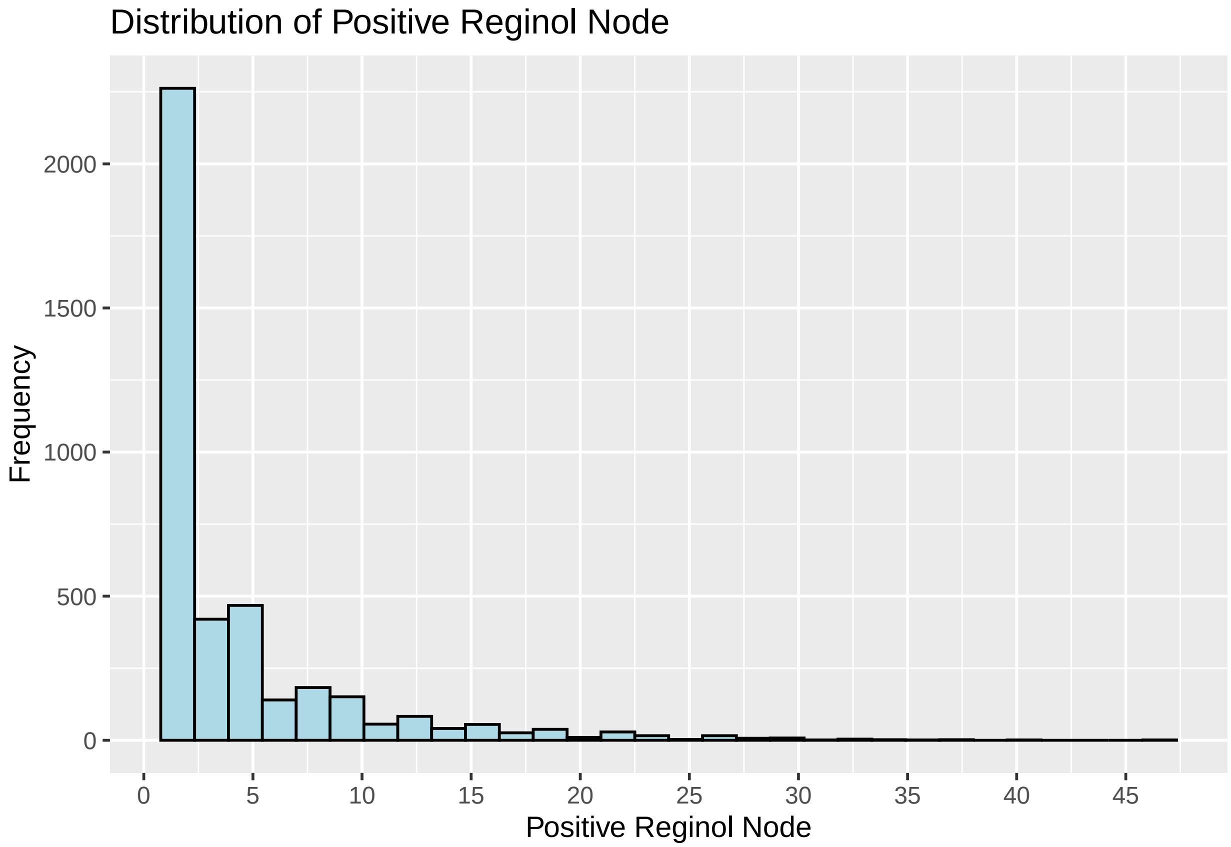
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



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.

```
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"
  )
```

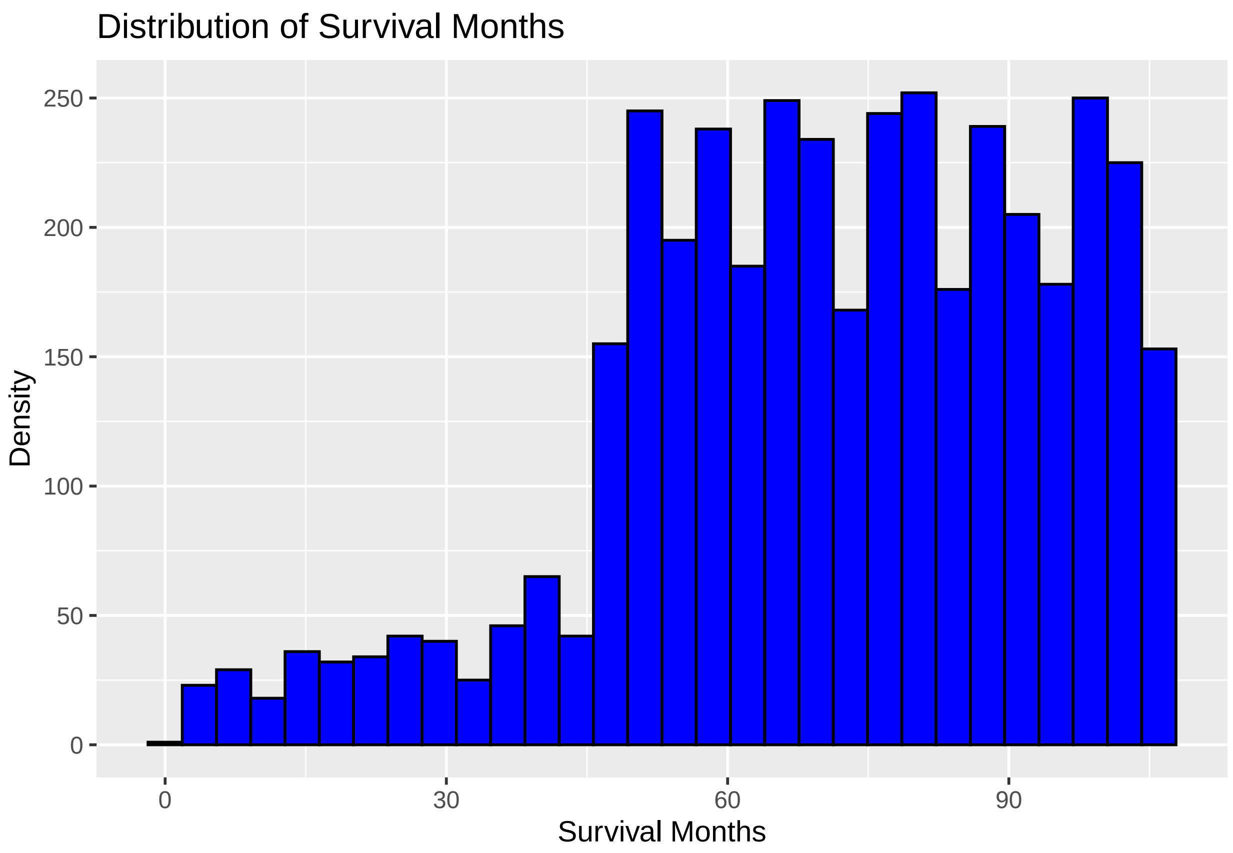
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



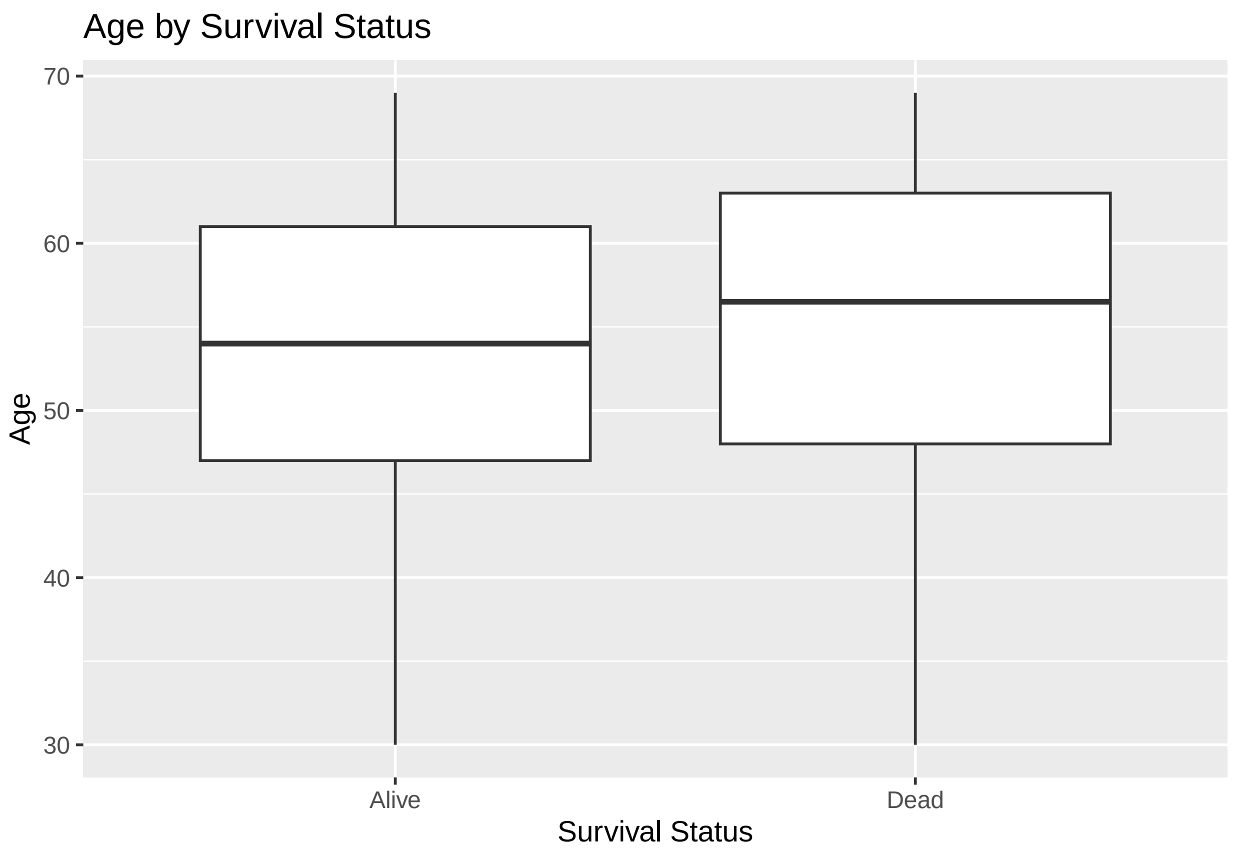
Then is the distribution of different number of positive reginol node for each subject. Over 2500 subjects only have 1 or 2 positive reginol nodes, which is the most frequent number of positive reginol nodes.

```
ggplot(survival_df, aes(x = survival_months)) +
  geom_histogram(fill = "blue", color = "black") +
  labs(title = "Distribution of Survival Months", x = "Survival Months", y = "Density")
```

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



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

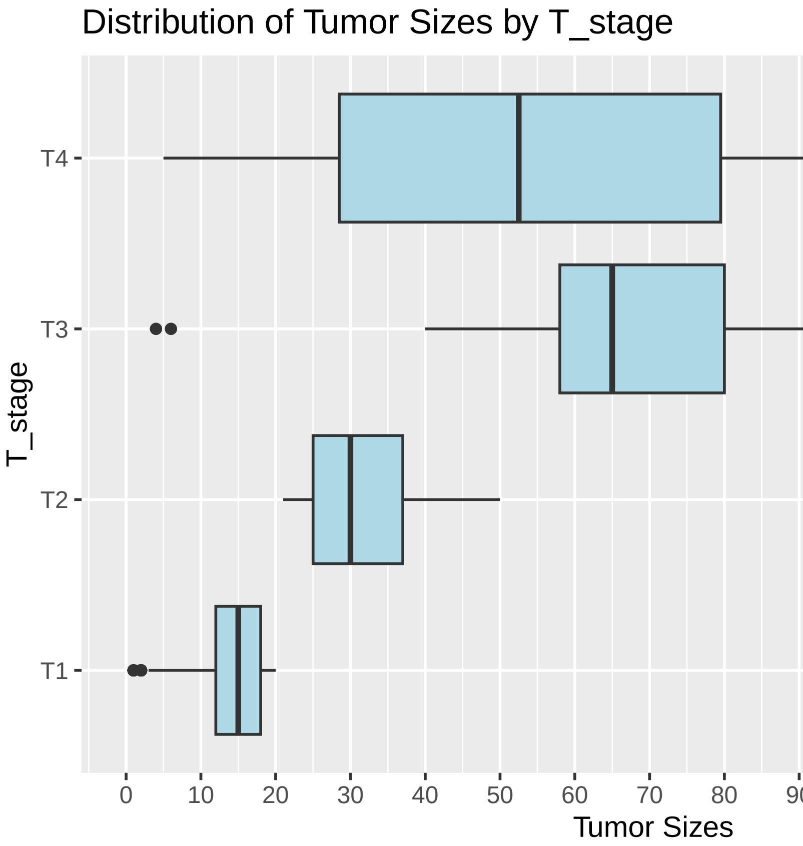


```
survival_df |>
  group_by(race) |>
  summarize(Count = n(), Proportion = n() / nrow(survival_df)) |>
  knitr::kable()
```

race	Count	Proportion
Black	291	0.0723161
Other	320	0.0795229
White	3413	0.8481610

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.

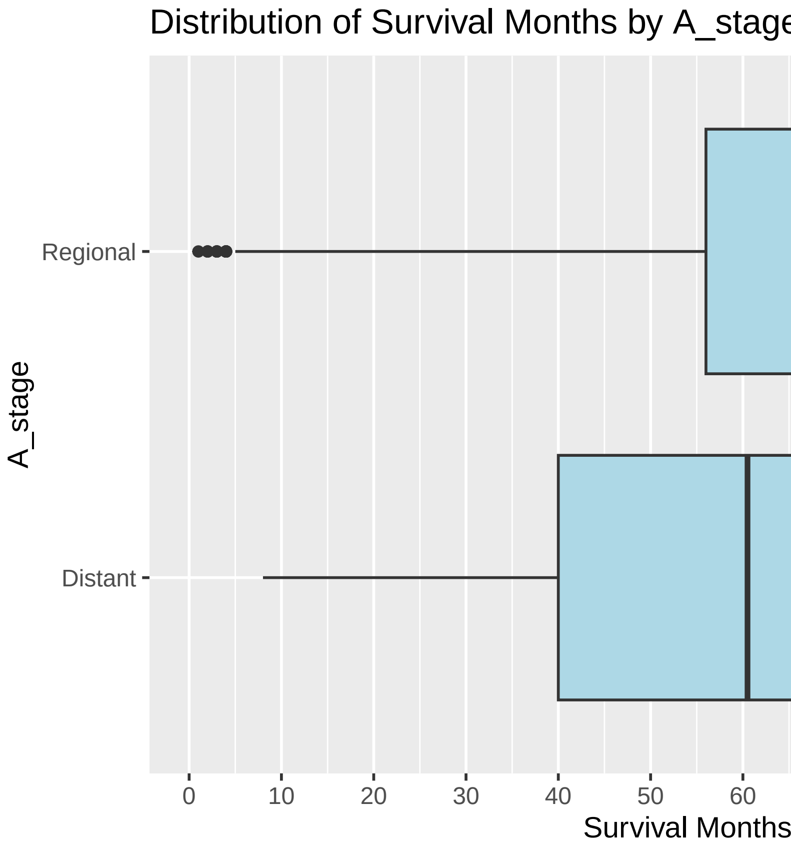
```
ggplot(survival_df, aes(x = tumor_size, y = t_stage)) +
  geom_boxplot(fill = "light blue") +
  scale_x_continuous(breaks = seq(0, max(survival_df$tumor_size, na.rm = TRUE), by = 10)) +
  labs(
    title = "Distribution of Tumor Sizes by T_stage",
    x = "Tumor Sizes",
    y = "T_stage"
  )
```



The distribution of the tumor sizes by 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.

```
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"
  )
```



The distribution of survival months by a__stage

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.