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" "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`.

These can be converted into factors for later analysis.

```
summary(survival_df)
```

```
##       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
##
##
##
```

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))
```

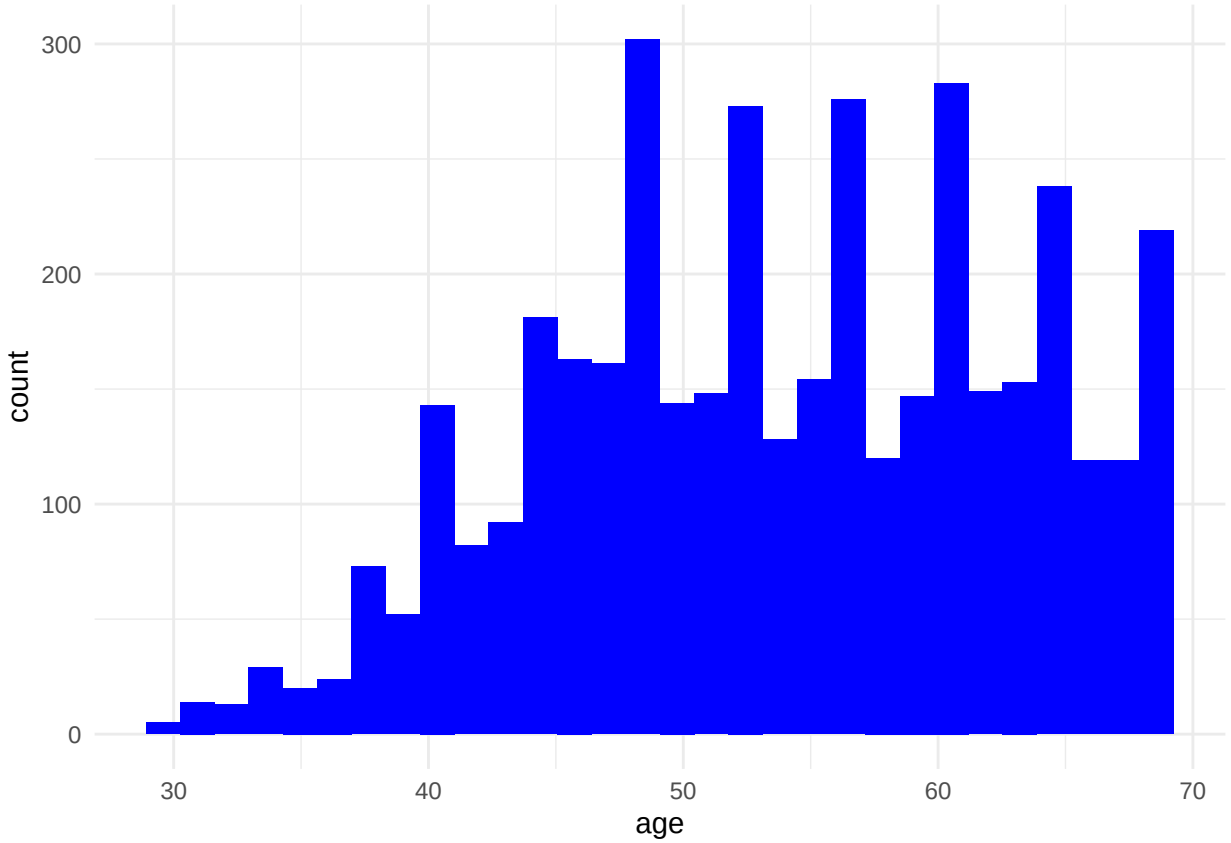
```
##                    age                   race         marital_status
##                      0                      0                      0
##                t_stage                n_stage             x6th_stage
##                      0                      0                      0
##          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") +
  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.