P8108 Final Project

Group 6

2025-11-16

Data Import

```
lung_df <- survival::lung %>%
  janitor::clean_names() %>%
  mutate(
    inst = as.factor(inst),
    time = as.numeric(time),
    status = as.factor(status),
    event = status == 2,
    age = as.numeric(age),
    sex = factor(sex, levels = c(1, 2), labels = c("Male", "Female")),
    ph_ecog = factor(ph_ecog, ordered = TRUE),
    ph_karno = as.numeric(ph_karno),
    pat_karno = as.numeric(pat_karno),
    meal_cal = as.numeric(meal_cal),
    wt_loss = as.numeric(wt_loss)
  )
```

We create a new variable called **event** to indicate survival status, where 1 represents death and 0 represents censoring.

The variable **ph_ecog** (ECOG performance score, 0–3) is treated as an ordinal variable. For descriptive and Kaplan–Meier analyses, it is handled as a categorical factor to visualize group differences. It might be modeled as an ordinal numeric variable in Cox model.

Check NAs

```
summary(lung_df)
```

```
##       inst          time        status       age            sex      ph_ecog
##  1      : 36   Min.   :   5.0   1: 63   Min.   :39.00   Male  :138   0   : 63
##  12     : 23   1st Qu.: 166.8   2:165   1st Qu.:56.00   Female: 90   1   :113
##  13     : 20   Median : 255.5           Median :63.00                2   : 50
##  3      : 19   Mean   : 305.2           Mean   :62.45                3   :  1
##  11     : 18   3rd Qu.: 396.5           3rd Qu.:69.00                NA's:  1
##  (Other):111   Max.   :1022.0           Max.   :82.00
##  NA's   :  1
##     ph_karno        pat_karno         meal_cal         wt_loss
##  Min.   : 50.00   Min.   : 30.00   Min.   :  96.0   Min.   :-24.000
##  1st Qu.: 75.00   1st Qu.: 70.00   1st Qu.: 635.0   1st Qu.:  0.000
##  Median : 80.00   Median : 80.00   Median : 975.0   Median :  7.000
##  Mean   : 81.94   Mean   : 79.96   Mean   : 928.8   Mean   :  9.832
##  3rd Qu.: 90.00   3rd Qu.: 90.00   3rd Qu.:1150.0   3rd Qu.: 15.750
##  Max.   :100.00   Max.   :100.00   Max.   :2600.0   Max.   : 68.000
##  NA's   :1        NA's   :3        NA's   :47       NA's   :14
##    event
##  Mode :logical
##  FALSE:63
##  TRUE :165
##
##
##
##
```

```
lung_cc <- lung_df %>%
  filter(complete.cases(time, event, age, sex, ph_ecog,
                        ph_karno, pat_karno, meal_cal, wt_loss, inst))
summary(lung_cc)
```

```
##       inst         time        status       age            sex      ph_ecog
##  1      :28   Min.   :   5.0   1: 47   Min.   :39.00   Male  :103   0:47
##  12     :16   1st Qu.: 174.5   2:120   1st Qu.:57.00   Female: 64   1:81
##  11     :13   Median : 268.0           Median :64.00                2:38
##  13     :13   Mean   : 309.9           Mean   :62.57                3: 1
##  22     :13   3rd Qu.: 419.5           3rd Qu.:70.00
##  3      :12   Max.   :1022.0           Max.   :82.00
##  (Other):72
##     ph_karno        pat_karno         meal_cal         wt_loss
##  Min.   : 50.00   Min.   : 30.00   Min.   :  96.0   Min.   :-24.000
##  1st Qu.: 70.00   1st Qu.: 70.00   1st Qu.: 619.0   1st Qu.:  0.000
##  Median : 80.00   Median : 80.00   Median : 975.0   Median :  7.000
##  Mean   : 82.04   Mean   : 79.58   Mean   : 929.1   Mean   :  9.719
##  3rd Qu.: 90.00   3rd Qu.: 90.00   3rd Qu.:1162.5   3rd Qu.: 15.000
##  Max.   :100.00   Max.   :100.00   Max.   :2600.0   Max.   : 68.000
##
##    event
##  Mode :logical
##  FALSE:47
##  TRUE :120
##
##
##
##
```

There are NAs in this data, so we will also create another dataset that observations with missing values will be excluded to ensure that all variables used in the analysis had complete information.

Both the original dataset and the complete-case dataset were retained for further analyses to allow comparisons and sensitivity checks.

EDA

Descriptive Table

```
lung_cc %>%
  select(time, event, age, sex, ph_ecog, ph_karno, pat_karno, meal_cal, wt_loss) %>%
  tbl_summary(
    by = sex,
    missing = "no",
```

Characteristic	**Overall** N = 167[1]	**Male** N = 103[1]	**Female** N = 64[1]	**p-value**[2]
Survival time (days)	310 (209)	291 (208)	340 (209)	0.069
Death indicator	120 (72%)	82 (80%)	38 (59%)	0.005
Age (years)	63 (9)	63 (9)	61 (9)	0.10
ECOG performance status				>0.9
0	47 (28%)	28 (27%)	19 (30%)	
1	81 (49%)	52 (50%)	29 (45%)	
2	38 (23%)	22 (21%)	16 (25%)	
3	1 (0.6%)	1 (1.0%)	0 (0%)	
Physician Karnofsky score				0.8
50	4 (2.4%)	3 (2.9%)	1 (1.6%)	
60	16 (9.6%)	8 (7.8%)	8 (13%)	
70	24 (14%)	16 (16%)	8 (13%)	
80	47 (28%)	28 (27%)	19 (30%)	
90	50 (30%)	32 (31%)	18 (28%)	
100	26 (16%)	16 (16%)	10 (16%)	
Patient Karnofsky score				0.3
30	2 (1.2%)	1 (1.0%)	1 (1.6%)	
40	2 (1.2%)	1 (1.0%)	1 (1.6%)	
50	3 (1.8%)	2 (1.9%)	1 (1.6%)	
60	23 (14%)	15 (15%)	8 (13%)	
70	30 (18%)	24 (23%)	6 (9.4%)	
80	37 (22%)	20 (19%)	17 (27%)	
90	44 (26%)	24 (23%)	20 (31%)	
100	26 (16%)	16 (16%)	10 (16%)	
Meal calories	929 (413)	985 (428)	840 (374)	0.027
Weight loss (kg)	10 (13)	12 (13)	7 (13)	0.015

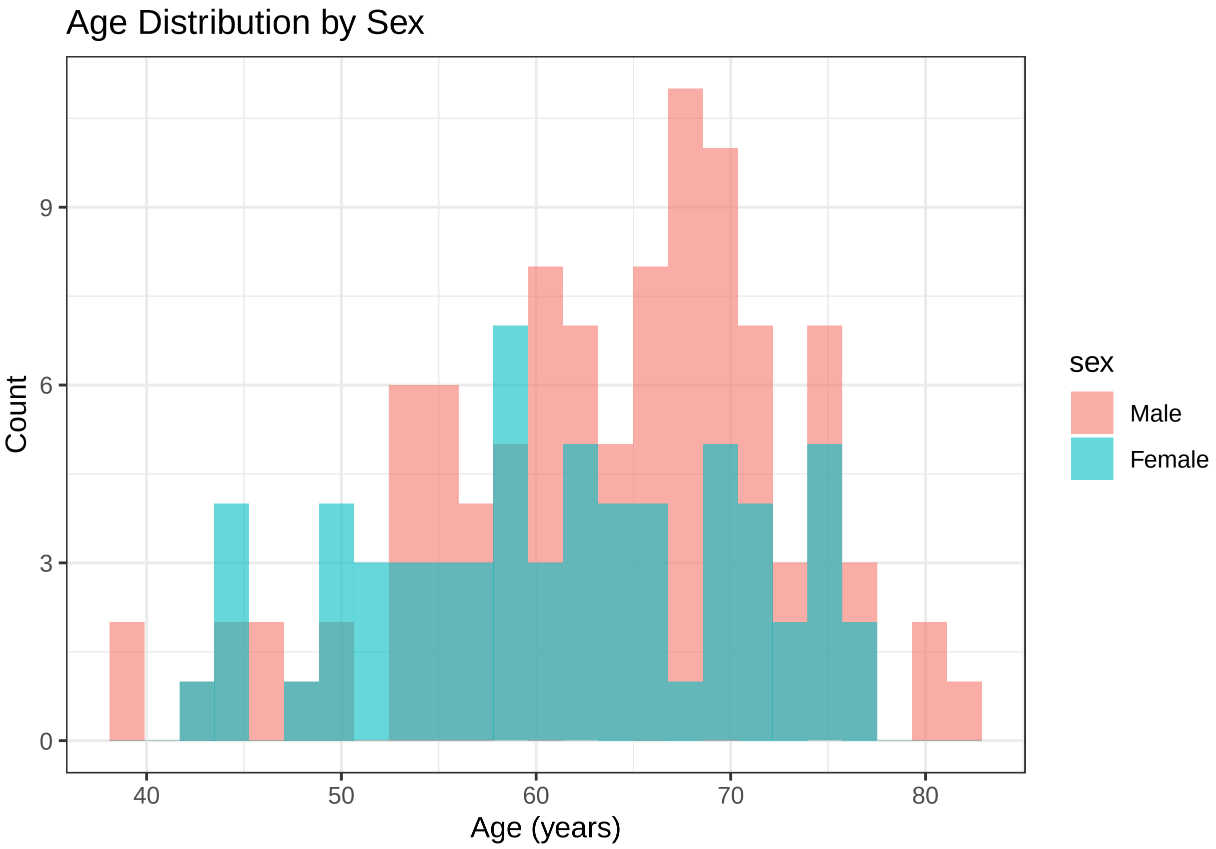
[1] Mean (SD); n (%)

[2] Wilcoxon rank sum test

```r
    statistic = list(all_continuous() ~ "{mean} ({sd})", all_categorical() ~ "{n} ({p}%)"),
    label = list(
      time ~ "Survival time (days)",
      event ~ "Death indicator",
      age ~ "Age (years)",
      ph_ecog ~ "ECOG performance status",
      ph_karno ~ "Physician Karnofsky score",
      pat_karno ~ "Patient Karnofsky score",
      meal_cal ~ "Meal calories",
      wt_loss ~ "Weight loss (kg)"
    )
  ) %>%
  add_overall() %>%
  add_p(test = everything() ~ "wilcox.test") %>%
  bold_labels()
```

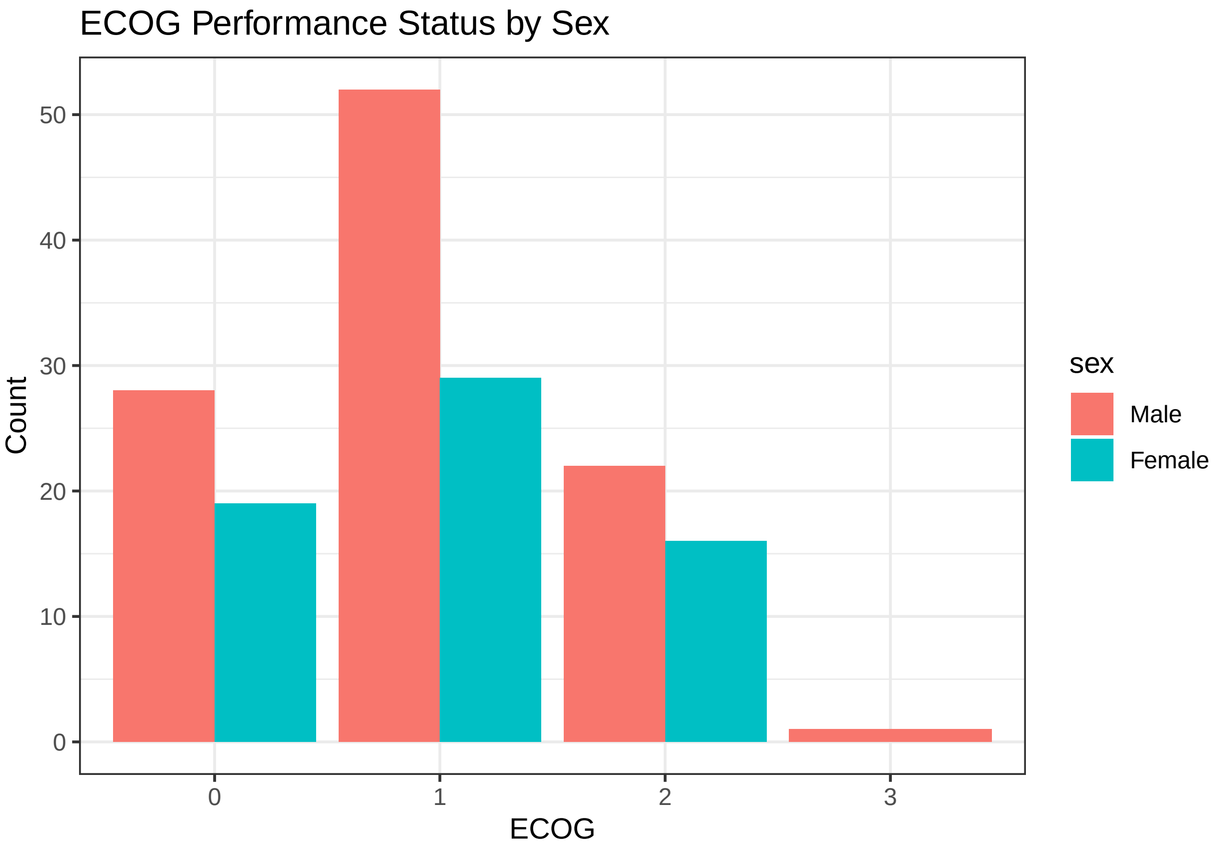
Age Distribution by Sex

```r
ggplot(lung_cc, aes(x = age, fill = sex)) +
  geom_histogram(bins = 25, position = "identity", alpha = 0.6) +
  theme_bw() +
  labs(title = "Age Distribution by Sex", x = "Age (years)", y = "Count")
```



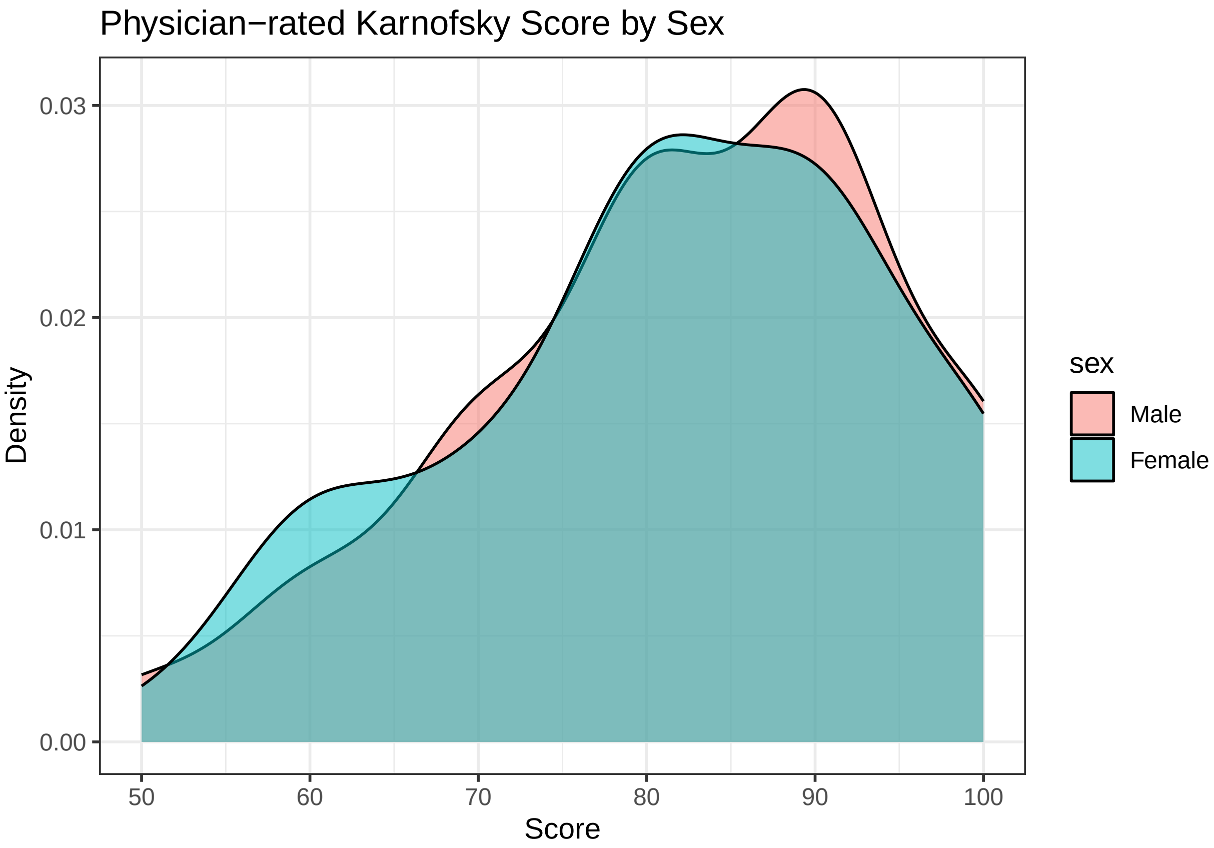
ECOG Performance Status by Sex

```r
ggplot(lung_cc, aes(x = ph_ecog, fill = sex)) +
  geom_bar(position = "dodge") +
  theme_bw() +
  labs(title = "ECOG Performance Status by Sex", x = "ECOG", y = "Count")
```



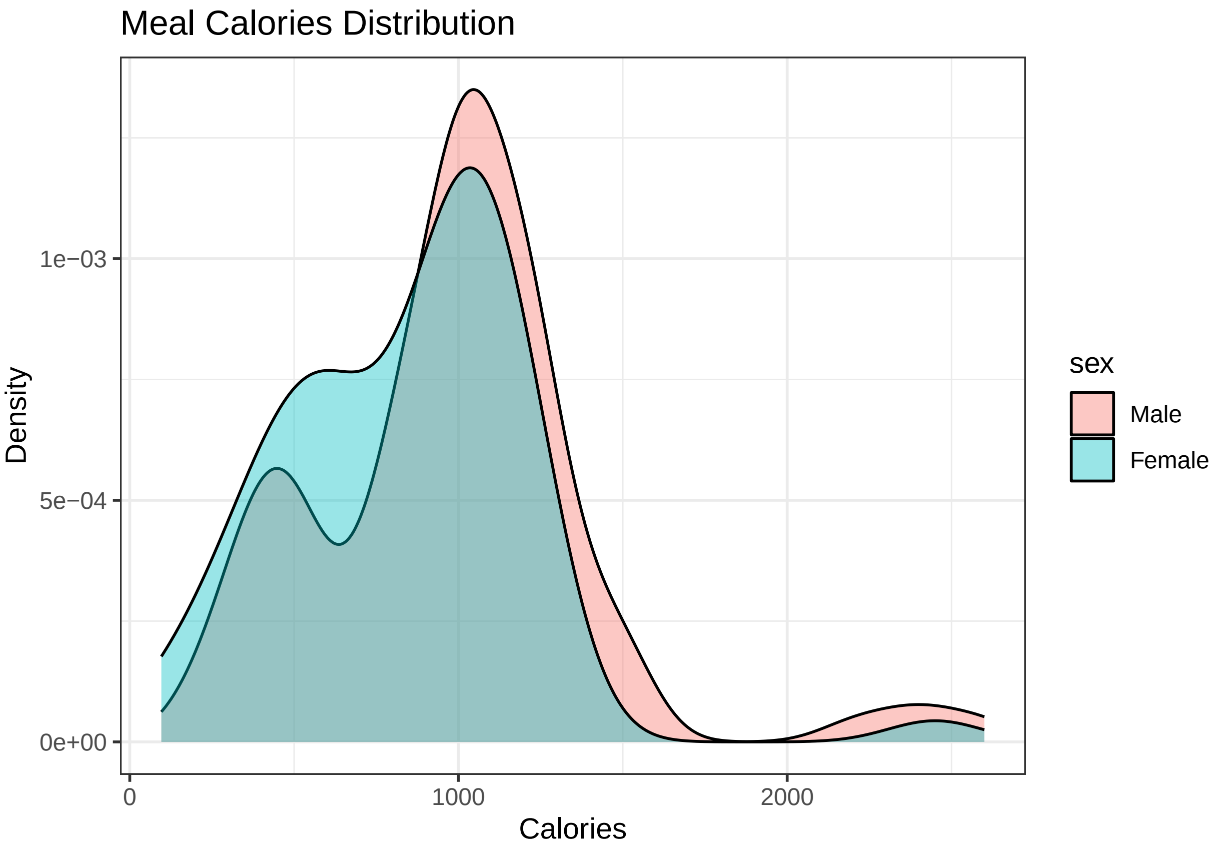
Physician-rated Karnofsky Score by Sex

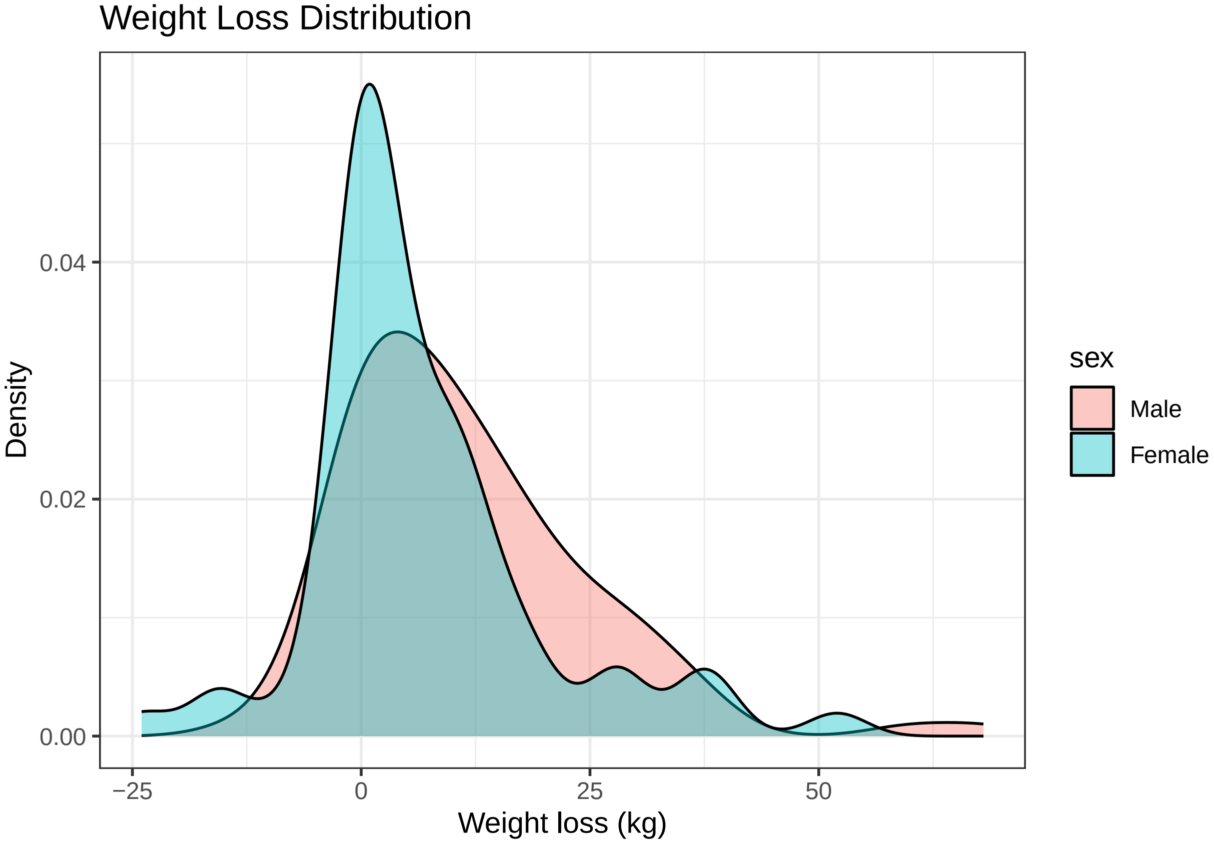
```
ggplot(lung_cc, aes(x = ph_karno, fill = sex)) +
  geom_density(alpha = 0.5) +
  theme_bw() +
  labs(title = "Physician-rated Karnofsky Score by Sex", x = "Score", y = "Density")
```



Meal Calories and Weight Loss Distributions

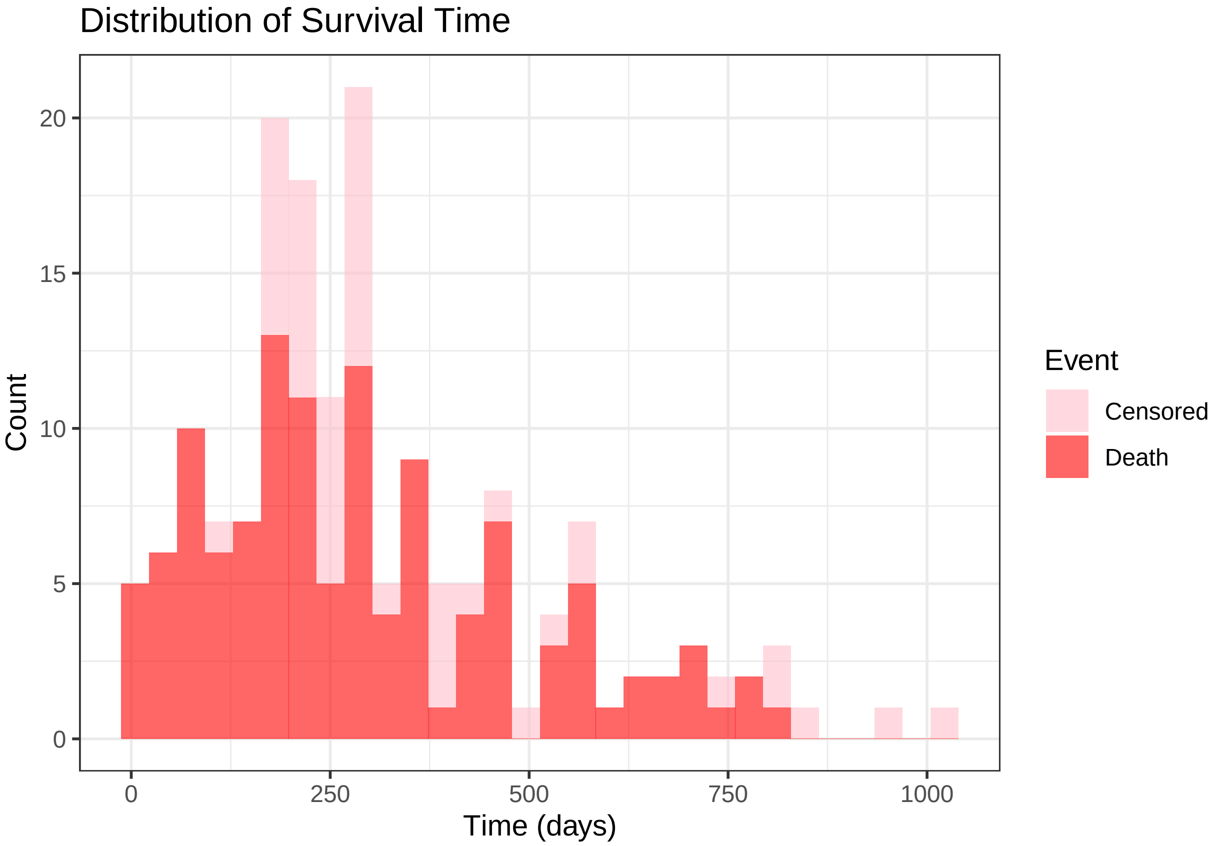
```r
p1 <- ggplot(lung_cc, aes(x = meal_cal, fill = sex)) +
  geom_density(alpha = 0.4) +
  labs(title = "Meal Calories Distribution", x = "Calories", y = "Density") +
  theme_bw()

p2 <- ggplot(lung_cc, aes(x = wt_loss, fill = sex)) +
  geom_density(alpha = 0.4) +
  labs(title = "Weight Loss Distribution", x = "Weight loss (kg)", y = "Density") +
  theme_bw()

p1; p2
```


Meal Calories Distribution
Density
1e−03
5e−04
0e+00
0
1000
2000
Calories
sex
Male
Female


Weight Loss Distribution
Density
0.04
0.02
0.00
−25
0
25
50
Weight loss (kg)
sex
Male
Female

Distribution of Survival Time

```r
ggplot(lung_cc, aes(x = time, fill = event)) +
  geom_histogram(bins = 30, alpha = 0.6) +
  theme_bw() +
  labs(title = "Distribution of Survival Time", x = "Time (days)", y = "Count") +
  scale_fill_manual(values = c("PINK", "RED"), name = "Event", labels = c("Censored", "Death"))
```


Distribution of Survival Time
Count
Time (days)
Event
Censored
Death