

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",
    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()
```

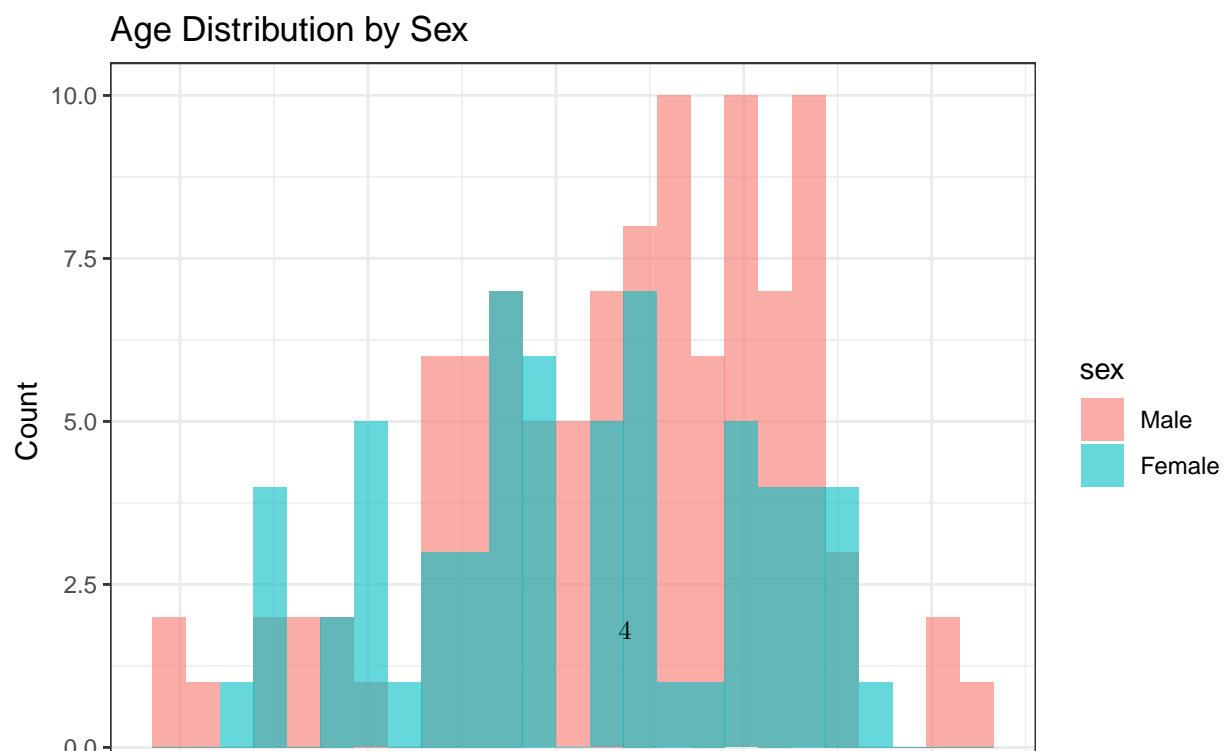
```
## The following errors were returned during 'add_p()':
## x For variable 'age' ('sex') and "p.value" statistic: The package "broom" (>=
## 1.0.8) is required.
## x For variable 'event' ('sex') and "p.value" statistic: The package "broom" (>=
## 1.0.8) is required.
## x For variable 'meal_cal' ('sex') and "p.value" statistic: The package "broom"
## (>= 1.0.8) is required.
## x For variable 'pat_karno' ('sex') and "p.value" statistic: The package "broom"
## (>= 1.0.8) is required.
## x For variable 'ph_ecog' ('sex') and "p.value" statistic: The package "broom"
## (>= 1.0.8) is required.
## x For variable 'ph_karno' ('sex') and "p.value" statistic: The package "broom"
## (>= 1.0.8) is required.
## x For variable 'time' ('sex') and "p.value" statistic: The package "broom" (>=
## 1.0.8) is required.
## x For variable 'wt_loss' ('sex') and "p.value" statistic: The package "broom"
## (>= 1.0.8) is required.
```

Age Distribution by Sex

```
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")
```

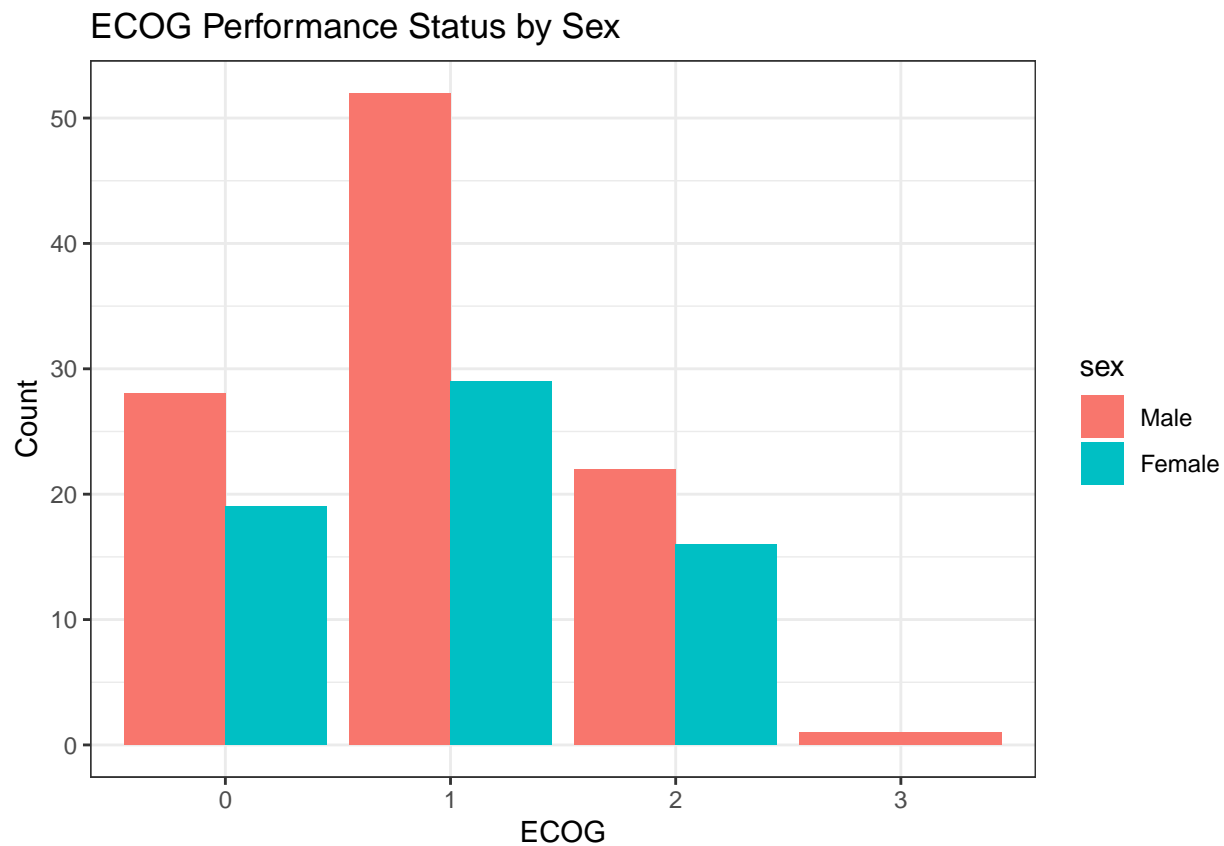
Characteristic	Overall N = 167 [†]	Male N = 103 [†]	Female N = 64 [†]	p-value
Survival time (days)	310 (209)	291 (208)	340 (209)	
Death indicator	120 (72%)	82 (80%)	38 (59%)	
Age (years)	63 (9)	63 (9)	61 (9)	
ECOG performance status				
0	47 (28%)	28 (27%)	19 (30%)	
1	81 (49%)	52 (50%)	29 (45%)	
2	38 (23%)	22 (21%)	16 (25%)	
3	1 (0.6%)	1 (1.0%)	0 (0%)	
Physician Karnofsky score				
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				
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)	
Weight loss (kg)	10 (13)	12 (13)	7 (13)	

[†]Mean (SD); n (%)



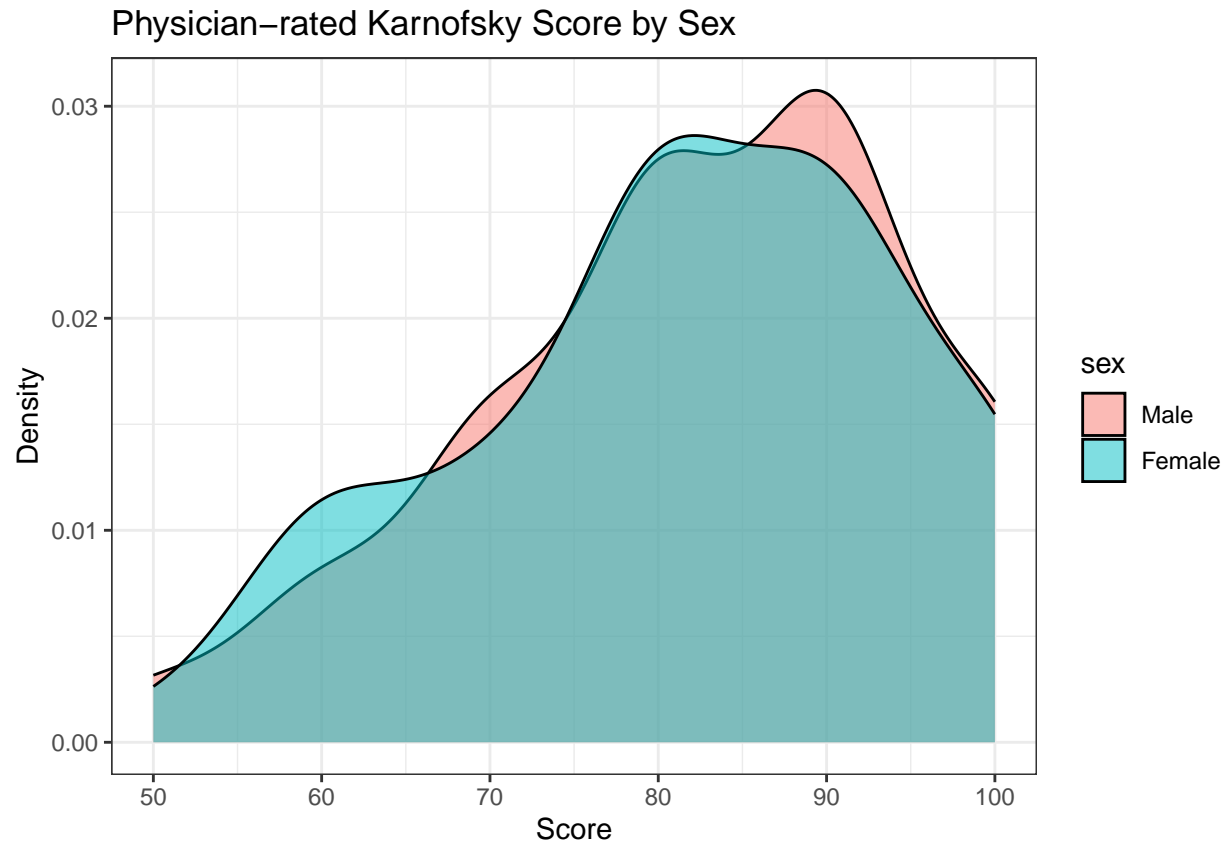
ECOG Performance Status by Sex

```
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

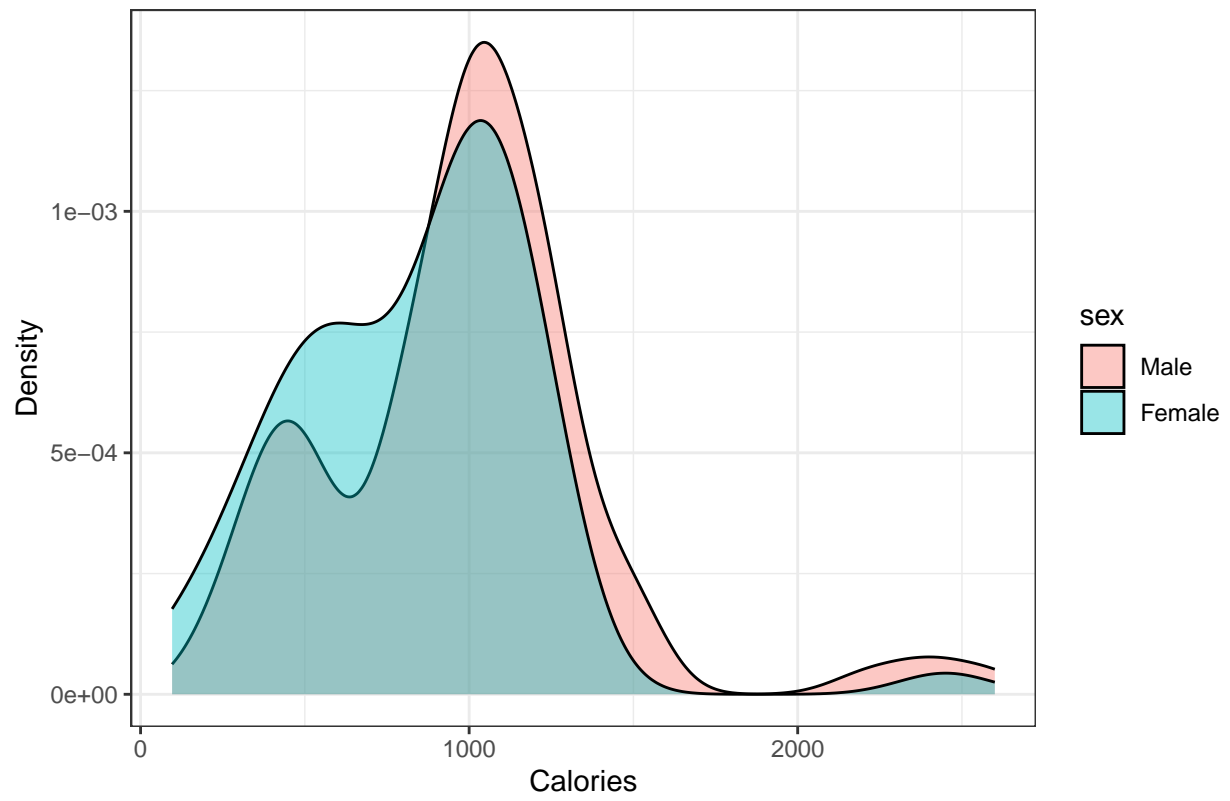
```
ggplot(lung_cc, aes(x = ph_karno, fill = sex)) +  
  geom_density(alpha = 0.5) +  
  theme_bw() +  
  labs(title = "Physician-rated Karnofsky Score by Sex", x = "Score", y = "Density")
```



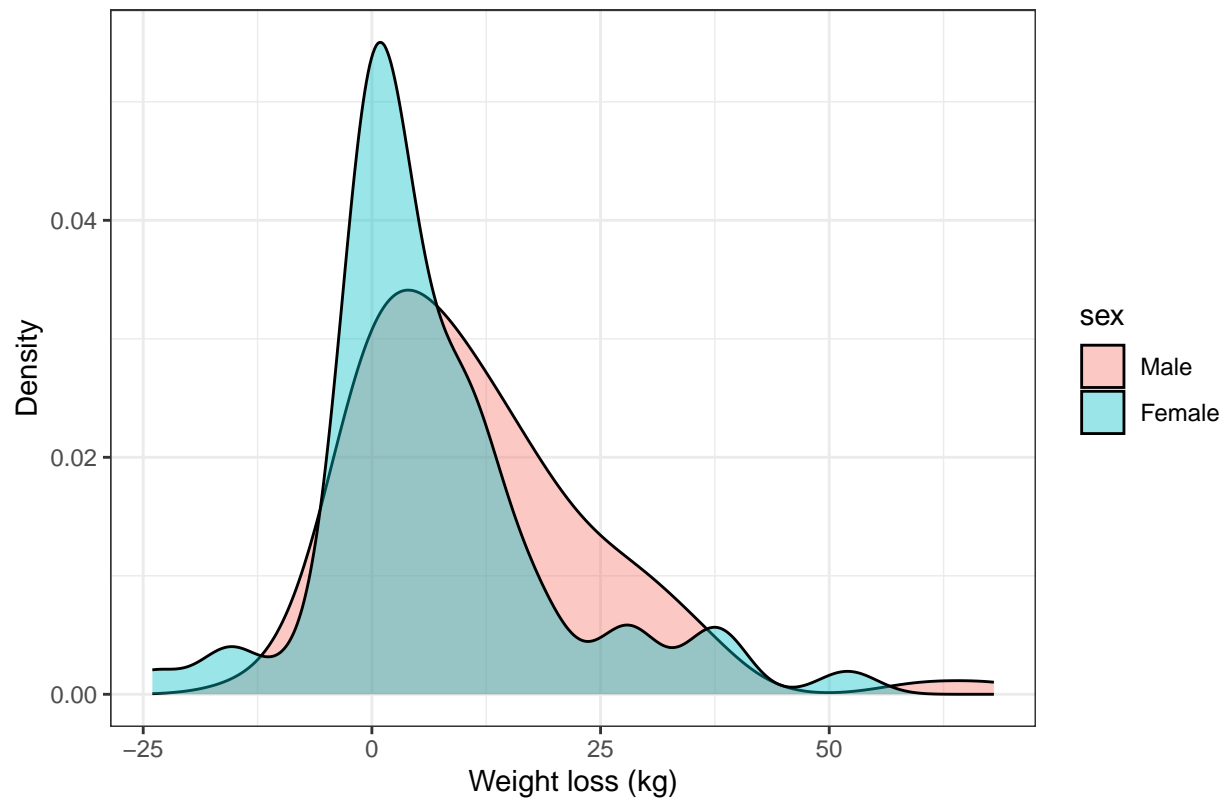
Meal Calories and Weight Loss Distributions

```
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



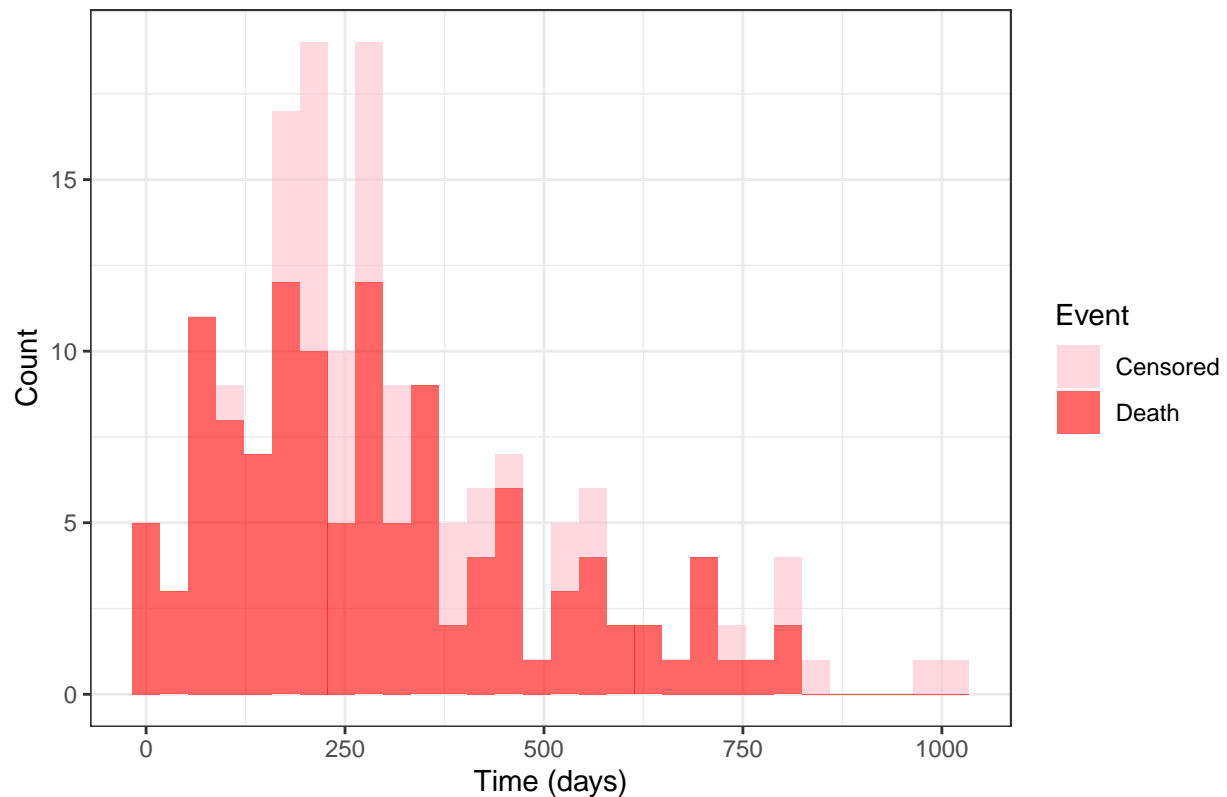
Weight Loss Distribution



Distribution of Survival Time

```
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



Comparison of Survival Curves

```
# Create a survival object, which bundles the time and event data
surv_object <- Surv(time = lung_cc$time, event = lung_cc$event)
```

Sex

```
# Fit the Kaplan-Meier model to estimate survival curves for each group ('sex')
km_sex_fit <- survfit(surv_object ~ sex, data = lung_cc)

summary(km_sex_fit)
```

```
## Call: survfit(formula = surv_object ~ sex, data = lung_cc)
##
##               sex=Male
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    11    103      1  0.9903 0.00966   0.9715      1.000
##    12    102      1  0.9806 0.01360   0.9543      1.000
##    13    101      1  0.9709 0.01657   0.9389      1.000
##    15    100      1  0.9612 0.01904   0.9246      0.999
##    26     99      1  0.9515 0.02118   0.9108      0.994
```

##	30	98	1	0.9417	0.02308	0.8976	0.988
##	31	97	1	0.9320	0.02480	0.8847	0.982
##	53	96	2	0.9126	0.02782	0.8597	0.969
##	54	94	1	0.9029	0.02917	0.8475	0.962
##	59	93	1	0.8932	0.03043	0.8355	0.955
##	60	92	1	0.8835	0.03161	0.8237	0.948
##	65	91	1	0.8738	0.03272	0.8119	0.940
##	88	90	1	0.8641	0.03377	0.8004	0.933
##	92	89	1	0.8544	0.03476	0.7889	0.925
##	93	88	1	0.8447	0.03569	0.7775	0.918
##	95	87	1	0.8350	0.03658	0.7663	0.910
##	110	86	1	0.8252	0.03742	0.7551	0.902
##	118	85	1	0.8155	0.03822	0.7440	0.894
##	135	84	1	0.8058	0.03898	0.7329	0.886
##	142	83	1	0.7961	0.03970	0.7220	0.878
##	147	82	1	0.7864	0.04038	0.7111	0.870
##	156	81	2	0.7670	0.04165	0.6895	0.853
##	163	79	3	0.7379	0.04333	0.6576	0.828
##	166	76	1	0.7282	0.04384	0.6471	0.819
##	170	75	1	0.7184	0.04432	0.6366	0.811
##	176	73	1	0.7086	0.04479	0.6260	0.802
##	179	72	1	0.6988	0.04523	0.6155	0.793
##	180	71	1	0.6889	0.04566	0.6050	0.784
##	181	70	2	0.6692	0.04642	0.5842	0.767
##	183	68	1	0.6594	0.04677	0.5738	0.758
##	197	64	1	0.6491	0.04716	0.5629	0.748
##	207	62	1	0.6386	0.04755	0.5519	0.739
##	210	61	1	0.6282	0.04791	0.5409	0.729
##	212	60	1	0.6177	0.04824	0.5300	0.720
##	218	59	1	0.6072	0.04855	0.5191	0.710
##	222	57	1	0.5966	0.04885	0.5081	0.700
##	223	55	1	0.5857	0.04915	0.4969	0.690
##	229	52	1	0.5745	0.04948	0.4852	0.680
##	230	51	1	0.5632	0.04977	0.4736	0.670
##	246	50	1	0.5519	0.05004	0.4621	0.659
##	267	48	1	0.5404	0.05030	0.4503	0.649
##	269	47	1	0.5289	0.05053	0.4386	0.638
##	270	46	1	0.5174	0.05072	0.4270	0.627
##	283	45	1	0.5059	0.05088	0.4154	0.616
##	284	44	1	0.4944	0.05101	0.4039	0.605
##	285	42	1	0.4827	0.05113	0.3922	0.594
##	286	41	1	0.4709	0.05122	0.3805	0.583
##	288	40	1	0.4591	0.05128	0.3689	0.571
##	291	39	1	0.4473	0.05129	0.3573	0.560
##	301	36	1	0.4349	0.05135	0.3451	0.548
##	303	34	1	0.4221	0.05141	0.3325	0.536
##	320	32	1	0.4089	0.05147	0.3195	0.523
##	337	31	1	0.3957	0.05147	0.3067	0.511
##	353	30	2	0.3694	0.05131	0.2813	0.485
##	363	28	1	0.3562	0.05114	0.2688	0.472
##	371	27	1	0.3430	0.05092	0.2564	0.459
##	390	26	1	0.3298	0.05064	0.2441	0.446
##	428	23	1	0.3154	0.05043	0.2306	0.432
##	429	22	1	0.3011	0.05014	0.2173	0.417

##	455	21	1	0.2868	0.04976	0.2041	0.403
##	457	20	1	0.2724	0.04929	0.1911	0.388
##	460	18	1	0.2573	0.04882	0.1774	0.373
##	477	17	1	0.2422	0.04824	0.1639	0.358
##	519	16	1	0.2270	0.04754	0.1506	0.342
##	524	15	1	0.2119	0.04672	0.1375	0.326
##	558	14	1	0.1968	0.04577	0.1247	0.310
##	567	13	1	0.1816	0.04468	0.1121	0.294
##	574	12	1	0.1665	0.04344	0.0998	0.278
##	583	11	1	0.1514	0.04205	0.0878	0.261
##	613	10	1	0.1362	0.04048	0.0761	0.244
##	643	9	1	0.1211	0.03870	0.0647	0.227
##	655	8	1	0.1059	0.03671	0.0537	0.209
##	689	7	1	0.0908	0.03444	0.0432	0.191
##	707	6	1	0.0757	0.03185	0.0332	0.173
##	791	5	1	0.0605	0.02886	0.0238	0.154
##	814	3	1	0.0404	0.02533	0.0118	0.138
##							
##							
				sex=Female			
##	time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
##	5	64	1	0.984	0.0155	0.9545	1.000
##	60	63	1	0.969	0.0217	0.9270	1.000
##	61	62	1	0.953	0.0264	0.9027	1.000
##	62	61	1	0.938	0.0303	0.8800	0.999
##	79	60	1	0.922	0.0335	0.8584	0.990
##	81	59	1	0.906	0.0364	0.8376	0.981
##	95	58	1	0.891	0.0390	0.8174	0.970
##	107	56	1	0.875	0.0414	0.7972	0.960
##	145	55	1	0.859	0.0436	0.7774	0.949
##	153	54	1	0.843	0.0456	0.7581	0.937
##	167	53	1	0.827	0.0475	0.7390	0.925
##	199	50	1	0.810	0.0493	0.7194	0.913
##	201	49	1	0.794	0.0510	0.7000	0.900
##	226	45	1	0.776	0.0528	0.6794	0.887
##	239	43	1	0.758	0.0546	0.6584	0.873
##	245	40	1	0.739	0.0564	0.6366	0.859
##	268	37	1	0.719	0.0583	0.6136	0.843
##	285	34	1	0.698	0.0603	0.5894	0.827
##	293	32	1	0.676	0.0623	0.5647	0.810
##	305	30	1	0.654	0.0641	0.5394	0.792
##	310	29	1	0.631	0.0658	0.5146	0.774
##	345	27	1	0.608	0.0674	0.4892	0.755
##	348	26	1	0.584	0.0687	0.4642	0.736
##	351	25	1	0.561	0.0698	0.4397	0.716
##	361	24	1	0.538	0.0707	0.4155	0.696
##	363	23	1	0.514	0.0714	0.3918	0.675
##	426	19	1	0.487	0.0726	0.3639	0.653
##	433	18	1	0.460	0.0734	0.3366	0.629
##	444	17	1	0.433	0.0739	0.3100	0.605
##	450	16	1	0.406	0.0741	0.2839	0.581
##	473	15	1	0.379	0.0739	0.2585	0.556
##	520	13	1	0.350	0.0738	0.2314	0.529
##	550	11	1	0.318	0.0736	0.2020	0.501
##	641	8	1	0.278	0.0744	0.1648	0.470

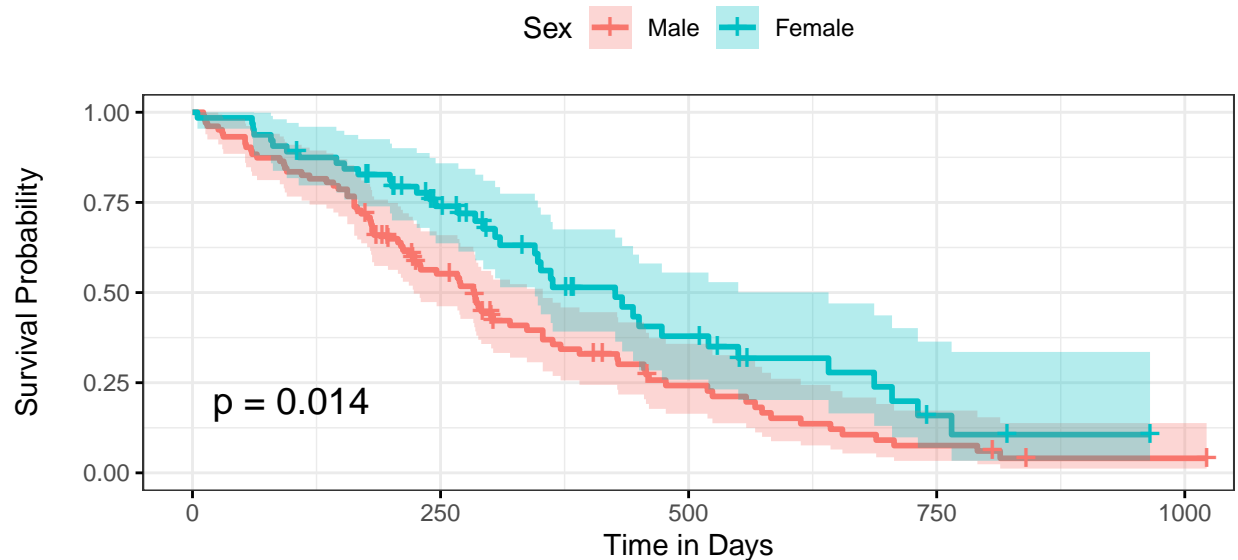
```
##    687      7      1    0.239 0.0736      0.1303      0.437
##    705      6      1    0.199 0.0713      0.0984      0.401
##    731      5      1    0.159 0.0672      0.0695      0.364
##    765      3      1    0.106 0.0623      0.0335      0.335
```

```
# --- Visualize the Estimated Survival Probabilities ---
```

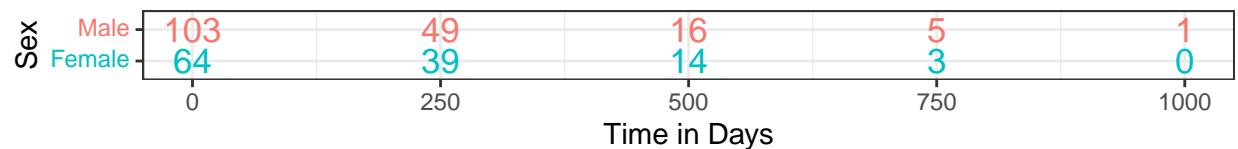
```
# Generate the Kaplan-Meier plot using ggsurvplot
```

```
ggsurvplot(
  km_sex_fit,
  data = lung_cc,
  pval = TRUE,                # The p-value from the log-rank test will be displayed
  conf.int = TRUE,           # Display confidence intervals for the curves
  risk.table = TRUE,         # Add a table showing the number of subjects at risk
  risk.table.col = "strata",  # Color the risk table to match the curves
  legend.labs = c("Male", "Female"),
  legend.title = "Sex",
  xlab = "Time in Days",
  ylab = "Survival Probability",
  title = "Kaplan-Meier Survival Curves by Sex",
  ggtheme = theme_bw()       # Apply a clean theme
)
```

Kaplan-Meier Survival Curves by Sex



Number at risk



```
# With rho = 0 this is the log-rank or Mantel-Haenszel test, and with rho = 1 it is equivalent to the P
log_rank_sex <- survdiff(surv_object ~ sex, data = lung_cc, rho = 0)
print(log_rank_sex)
```

```
## Call:
## survdiff(formula = surv_object ~ sex, data = lung_cc, rho = 0)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Male   103      82      68.7      2.57      6.05
## sex=Female  64      38      51.3      3.44      6.05
##
## Chisq= 6 on 1 degrees of freedom, p= 0.01
```

```
wilcoxon_sex <- survdiff(surv_object ~ sex, data = lung_cc, rho = 1)
print(wilcoxon_sex)
```

```
## Call:
## survdiff(formula = surv_object ~ sex, data = lung_cc, rho = 1)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=Male   103     51.0     41.8      2.01      6.76
## sex=Female  64     21.1     30.3      2.77      6.76
##
## Chisq= 6.8 on 1 degrees of freedom, p= 0.009
```

The p-value is even smaller than the log-rank p-value. This not only confirms the result of the log-rank test but also suggests that the survival advantage for females is particularly pronounced at the beginning and middle of the follow-up period.

```
# Fleming-Harrington test statistic
fit_sex <- ten(surv_object ~ sex, data = lung_cc)
comp(fit_sex)
```

```
##           Q          Var      Z pNorm
## 1          -13.2826    29.2085 -2.4577     5
## n        -1424.0000 303810.1223 -2.5835     4
## sqrtN      -133.2089   2591.7008 -2.6166     1
## S1          -9.0659    12.2163 -2.5938     2
## S2          -8.9740    11.9825 -2.5925     3
## FH_p=1_q=1   -2.1759     1.0197 -2.1548     6
##           maxAbsZ      Var      Q pSupBr
## 1          1.3806e+01 2.9209e+01 2.5545     5
## n          1.4410e+03 3.0381e+05 2.6143     4
## sqrtN       1.3534e+02 2.5917e+03 2.6585     1
## S1          9.1529e+00 1.2216e+01 2.6187     2
## S2          9.0560e+00 1.1982e+01 2.6162     3
## FH_p=1_q=1  2.2505e+00 1.0197e+00 2.2287     6
```

```
lrt_mat <- attr(fit_sex, "lrt")
data.frame(
  test = c("Log-Rank", "Wilcoxon", "Tarone", "Peto", "Modified-Peto", "FH(1, 1)"),
  Z_squared = lrt_mat[, "Z"]^2
)
```

```
##           test      Z
## 1      Log-Rank 6.040270
```

```
## 2      Wilcoxon 6.674485
## 3      Tarone 6.846710
## 4      Peto 6.727883
## 5 Modified-Peto 6.720914
## 6      FH(1, 1) 4.643096
```

The final summary table (Z_squared) shows that regardless of the specific test used, the result is always highly significant (all Z^2 values correspond to p-values well below 0.05).

ECOG performance status (unstratified)

```
# --- Fit the Kaplan-Meier model for ph_ecog ---
km_ecog_fit <- survfit(surv_object ~ ph_ecog, data = lung_cc)

summary(km_ecog_fit)
```

```
## Call: survfit(formula = surv_object ~ ph_ecog, data = lung_cc)
##
##                ph_ecog=0
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    5      47      1   0.979  0.0210   0.9383      1.000
##   15      46      1   0.957  0.0294   0.9014      1.000
##   31      45      1   0.936  0.0357   0.8688      1.000
##   53      44      1   0.915  0.0407   0.8385      0.998
##   81      43      1   0.894  0.0450   0.8097      0.986
##  147      42      1   0.872  0.0487   0.7820      0.973
##  176      40      1   0.851  0.0521   0.7543      0.959
##  246      34      1   0.826  0.0563   0.7223      0.944
##  267      30      1   0.798  0.0607   0.6874      0.926
##  285      26      1   0.767  0.0657   0.6487      0.908
##  286      25      1   0.737  0.0699   0.6116      0.887
##  303      22      1   0.703  0.0743   0.5716      0.865
##  320      21      1   0.670  0.0779   0.5331      0.841
##  337      19      1   0.634  0.0814   0.4933      0.816
##  348      18      1   0.599  0.0842   0.4549      0.789
##  353      17      2   0.529  0.0878   0.3818      0.732
##  371      15      1   0.493  0.0887   0.3468      0.702
##  428      12      1   0.452  0.0904   0.3057      0.669
##  433      11      1   0.411  0.0910   0.2664      0.635
##  455      10      1   0.370  0.0907   0.2289      0.598
##  558       9      1   0.329  0.0895   0.1930      0.561
##  574       7      1   0.282  0.0882   0.1527      0.520
##  643       6      1   0.235  0.0851   0.1155      0.478
##  655       5      1   0.188  0.0800   0.0816      0.433
##  705       4      1   0.141  0.0725   0.0515      0.386
##  791       3      1   0.094  0.0617   0.0260      0.340
##
##                ph_ecog=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    59      81      1   0.9877  0.0123   0.9639      1.000
##    60      80      2   0.9630  0.0210   0.9227      1.000
##    62      78      1   0.9506  0.0241   0.9046      0.999
```

##	79	77	1	0.9383	0.0267	0.8873	0.992
##	88	76	1	0.9259	0.0291	0.8706	0.985
##	92	75	1	0.9136	0.0312	0.8544	0.977
##	95	74	1	0.9012	0.0331	0.8385	0.969
##	110	73	1	0.8889	0.0349	0.8230	0.960
##	135	72	1	0.8765	0.0366	0.8078	0.951
##	142	71	1	0.8642	0.0381	0.7927	0.942
##	145	70	1	0.8519	0.0395	0.7779	0.933
##	156	69	1	0.8395	0.0408	0.7633	0.923
##	163	68	2	0.8148	0.0432	0.7345	0.904
##	167	66	1	0.8025	0.0442	0.7203	0.894
##	170	65	1	0.7901	0.0452	0.7062	0.884
##	179	62	1	0.7774	0.0463	0.6918	0.874
##	181	61	2	0.7519	0.0481	0.6632	0.852
##	197	57	1	0.7387	0.0491	0.6485	0.841
##	207	54	1	0.7250	0.0500	0.6333	0.830
##	210	53	1	0.7113	0.0509	0.6182	0.818
##	218	52	1	0.6977	0.0517	0.6033	0.807
##	223	49	1	0.6834	0.0526	0.5877	0.795
##	226	47	1	0.6689	0.0535	0.5719	0.782
##	229	46	1	0.6543	0.0542	0.5562	0.770
##	230	45	1	0.6398	0.0550	0.5407	0.757
##	245	43	1	0.6249	0.0557	0.5248	0.744
##	268	42	1	0.6100	0.0563	0.5091	0.731
##	269	41	1	0.5952	0.0568	0.4936	0.718
##	270	40	1	0.5803	0.0573	0.4781	0.704
##	283	39	1	0.5654	0.0578	0.4628	0.691
##	284	38	1	0.5505	0.0581	0.4476	0.677
##	293	37	1	0.5356	0.0584	0.4325	0.663
##	301	35	1	0.5203	0.0587	0.4171	0.649
##	305	32	1	0.5041	0.0591	0.4006	0.634
##	345	31	1	0.4878	0.0594	0.3843	0.619
##	363	30	2	0.4553	0.0597	0.3521	0.589
##	390	27	1	0.4384	0.0598	0.3355	0.573
##	426	24	1	0.4202	0.0601	0.3175	0.556
##	429	23	1	0.4019	0.0602	0.2997	0.539
##	450	22	1	0.3836	0.0601	0.2821	0.522
##	457	21	1	0.3654	0.0600	0.2648	0.504
##	460	19	1	0.3461	0.0598	0.2467	0.486
##	473	18	1	0.3269	0.0595	0.2288	0.467
##	477	17	1	0.3077	0.0590	0.2112	0.448
##	519	16	1	0.2884	0.0584	0.1940	0.429
##	520	15	1	0.2692	0.0576	0.1770	0.409
##	550	13	1	0.2485	0.0568	0.1588	0.389
##	567	12	1	0.2278	0.0557	0.1411	0.368
##	583	11	1	0.2071	0.0543	0.1238	0.346
##	613	10	1	0.1864	0.0527	0.1071	0.324
##	641	9	1	0.1657	0.0507	0.0909	0.302
##	687	8	1	0.1450	0.0484	0.0753	0.279
##	689	7	1	0.1243	0.0457	0.0604	0.256
##	731	6	1	0.1035	0.0425	0.0463	0.232
##	765	4	1	0.0777	0.0390	0.0290	0.208

##

ph_ecog=2

```
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
## 11 38 1 0.9737 0.0260 0.9241 1.000
## 12 37 1 0.9474 0.0362 0.8790 1.000
## 13 36 1 0.9211 0.0437 0.8392 1.000
## 26 35 1 0.8947 0.0498 0.8023 0.998
## 30 34 1 0.8684 0.0548 0.7673 0.983
## 53 33 1 0.8421 0.0592 0.7338 0.966
## 54 32 1 0.8158 0.0629 0.7014 0.949
## 61 31 1 0.7895 0.0661 0.6699 0.930
## 65 30 1 0.7632 0.0690 0.6393 0.911
## 93 29 1 0.7368 0.0714 0.6093 0.891
## 95 28 1 0.7105 0.0736 0.5800 0.870
## 107 26 1 0.6832 0.0756 0.5499 0.849
## 153 25 1 0.6559 0.0774 0.5204 0.827
## 156 24 1 0.6285 0.0789 0.4915 0.804
## 163 23 1 0.6012 0.0800 0.4632 0.780
## 166 22 1 0.5739 0.0809 0.4353 0.757
## 180 21 1 0.5466 0.0815 0.4080 0.732
## 183 20 1 0.5192 0.0819 0.3811 0.707
## 199 19 1 0.4919 0.0820 0.3547 0.682
## 201 18 1 0.4646 0.0819 0.3288 0.656
## 212 16 1 0.4355 0.0818 0.3014 0.629
## 222 15 1 0.4065 0.0813 0.2747 0.602
## 239 14 1 0.3775 0.0805 0.2485 0.573
## 285 13 1 0.3484 0.0794 0.2229 0.545
## 288 12 1 0.3194 0.0779 0.1980 0.515
## 291 11 1 0.2904 0.0760 0.1738 0.485
## 310 9 1 0.2581 0.0741 0.1470 0.453
## 351 8 1 0.2258 0.0715 0.1214 0.420
## 361 7 1 0.1936 0.0682 0.0970 0.386
## 444 6 1 0.1613 0.0640 0.0741 0.351
## 524 4 1 0.1210 0.0594 0.0462 0.317
## 707 2 1 0.0605 0.0521 0.0112 0.327
## 814 1 1 0.0000 NaN NA NA
##
##           ph_ecog=3
## time n.risk n.event survival std.err lower 95% CI
## 118 1 1 0 NaN NA
## upper 95% CI
## NA
```

```
ggsurvplot(
  km_ecog_fit,
  data = lung_cc,

  # --- Add Key Statistical Information ---
  pval = TRUE,           # Display the log-rank test p-value
  conf.int = TRUE,       # Show 95% confidence intervals

  # --- Add Contextual Information ---
  risk.table = TRUE,      # Include a table showing the number of patients at risk
  risk.table.y.text.col = TRUE, # Color the risk table text to match curves
  risk.table.y.text = FALSE, # Remove y-axis tick labels from the risk table
```

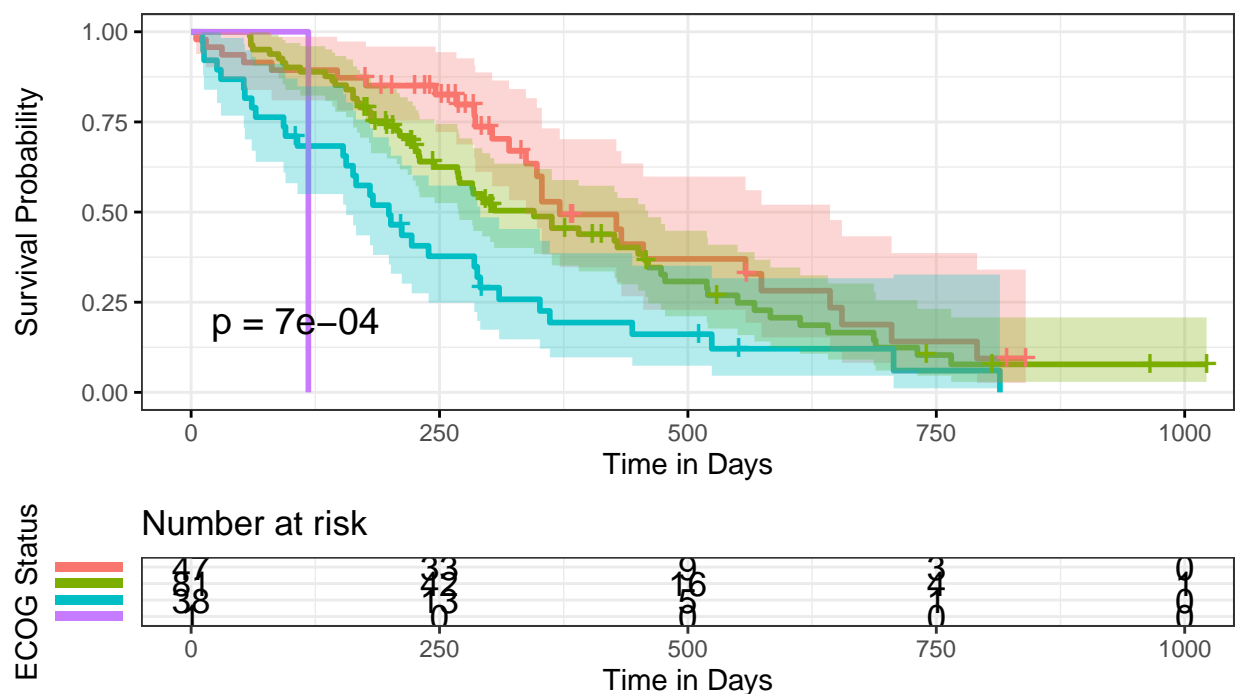
```

# --- Customize Aesthetics and Labels ---
title = "Kaplan-Meier Survival Curves by ECOG Performance Status",
xlab = "Time in Days",
ylab = "Survival Probability",
legend.title = "ECOG Status",
# Provide clear, descriptive labels for each group in the legend
legend.labs = c(
  "0 (Asymptomatic)",
  "1 (Symptomatic, Ambulatory)",
  "2 (In bed <50% of day)",
  "3 (In bed >50% of day)"
),
ggtheme = theme_bw() # Use a clean black and white theme
)

```

Kaplan-Meier Survival Curves by ECOG Performance Status

OG Status + 0 (Asymptomatic) + 1 (Symptomatic, Ambulatory) + 2 (In bed <50% of day) + 3 (In bed >50% of day)



```

# --- Perform the test ---
log_rank_ecog <- survdiff(surv_object ~ ph_ecog, data = lung_cc, rho = 0)
print(log_rank_ecog)

```

```

## Call:
## survdiff(formula = surv_object ~ ph_ecog, data = lung_cc, rho = 0)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## ph_ecog=0 47      27   38.054    3.211    4.737
## ph_ecog=1 81      59   62.376    0.183    0.383

```

```
## ph_ecog=2 38      33  19.394      9.546      11.483
## ph_ecog=3  1       1   0.176      3.864      3.895
##
## Chisq= 17  on 3 degrees of freedom, p= 7e-04
```

```
wilcoxon_ecog <- survdiff(surv_object ~ ph_ecog, data = lung_cc, rho = 1)
print(wilcoxon_ecog)
```

```
## Call:
## survdiff(formula = surv_object ~ ph_ecog, data = lung_cc, rho = 1)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## ph_ecog=0 47   14.231   22.518     3.050     6.299
## ph_ecog=1 81   33.801   37.054     0.286     0.826
## ph_ecog=2 38   23.250   12.392     9.513    15.668
## ph_ecog=3  1    0.844    0.162     2.878     3.144
##
## Chisq= 20.8  on 3 degrees of freedom, p= 1e-04
```

The fact that the Wilcoxon test yields a smaller p-value than the log-rank test suggests that the survival differences between the groups are particularly pronounced early in the follow-up period.

ECOG performance status (stratified by sex)

```
log_rank_ecog_bysex <- survdiff(surv_object ~ ph_ecog + strata(sex), data = lung_cc, rho = 0)
print(log_rank_ecog_bysex)
```

```
## Call:
## survdiff(formula = surv_object ~ ph_ecog + strata(sex), data = lung_cc,
##           rho = 0)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## ph_ecog=0 47      27   38.330     3.349     5.003
## ph_ecog=1 81      59   62.336     0.178     0.381
## ph_ecog=2 38      33   19.131    10.053    12.287
## ph_ecog=3  1       1    0.203     3.136     3.173
##
## Chisq= 17.1  on 3 degrees of freedom, p= 7e-04
```

```
wilcoxon_ecog_bysex <- survdiff(surv_object ~ ph_ecog + strata(sex), data = lung_cc, rho = 1)
print(wilcoxon_ecog_bysex)
```

```
## Call:
## survdiff(formula = surv_object ~ ph_ecog + strata(sex), data = lung_cc,
##           rho = 1)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## ph_ecog=0 47   13.478   22.324     3.505     7.26
## ph_ecog=1 81   34.002   37.497     0.326     0.96
## ph_ecog=2 38   23.686   11.987    11.419    18.70
```

```
## ph_ecog=3 1    0.825    0.184    2.226    2.47
##
##  Chisq= 23.4  on 3 degrees of freedom, p= 3e-05
```

Log-rank: The result is virtually identical to the unstratified test. The Chi-square statistic and the p-value have barely changed. It means that the strong predictive power of ECOG status is independent of the patient's sex. The effect is not being confounded by sex.

Wilcoxon: After stratifying by sex, the Chi-square value increased (from 20.8 to 23.4) and the p-value became even smaller. This suggests that after controlling for the baseline survival differences between sexes, the effect of ECOG status on early deaths becomes even more pronounced and clear.

Final summary for stratified analysis: ECOG is an Independent Predictor: The effect of ECOG performance status on survival is not explained away by the patient's sex. It is a strong, independent prognostic factor. Sex is not acting as a major confounder in the relationship between ECOG status and survival. Whether a patient is male or female, having a worse ECOG score (e.g., a score of 2 vs. 0) is strongly associated with a significantly poorer survival outcome. The effect holds true within both groups.