kaplan_meier_analysis

Import library

```r
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##      myeloma
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.6.0
## v lubridate 1.9.4     v tibble    3.3.0
## v purrr     1.2.0     v tidyr     1.3.2
```

```
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(readr)
library(dplyr)
library(gtsummary)
library(ggplot2)
```

Import the data

```r
data <- read.csv("~/Documents/Github/bio_stat_projects/data/seer_breat_cancer_dataset.csv", sep = ",")
```

Data Manipulation

```r
data$Race <- ifelse(data$Race == "Other (American Indian/AK Native, Asian/Pacific Islander)", "Other", 
data$Race <- as.factor(data$Race)

data$Marital.Status <- as.factor(data$Marital.Status)

data <- data %>%
  mutate(
    Marital.Status = case_when(
      Marital.Status == "Married (including common law)" ~ "Married",
      Marital.Status == "Single (never married)" ~ "Single",
      TRUE ~ Marital.Status # Keep everything else as it
    )
  )

data$T.Stage <- as.factor(data$T.Stage)
data$N.Stage <- as.factor(data$N.Stage)
data$X6th.Stage <- as.factor(data$X6th.Stage)
data$Grade <- as.factor(data$Grade)
data$A.Stage <- as.factor(data$A.Stage)
data$Estrogen.Status <- as.factor(data$Estrogen.Status)
data$Progesterone.Status <- as.factor(data$Progesterone.Status)
data$Status <- ifelse(data$Status == "Dead", 1, 0)

tbl_summary(data)
```

```r
library(survival)
library(survminer)

fit <- survfit(Surv(Survival.Months, Status) ~ 1, data = data)

# yes there is an empty plot but that disappears when you knit to pdf

ggsurvplot(
  fit,
  data = data,
  conf.int = TRUE,
  risk.table = TRUE,
  ggtheme = theme_minimal()
)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## i The deprecated feature was likely used in the ggpubr package.
##   Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.

## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
```
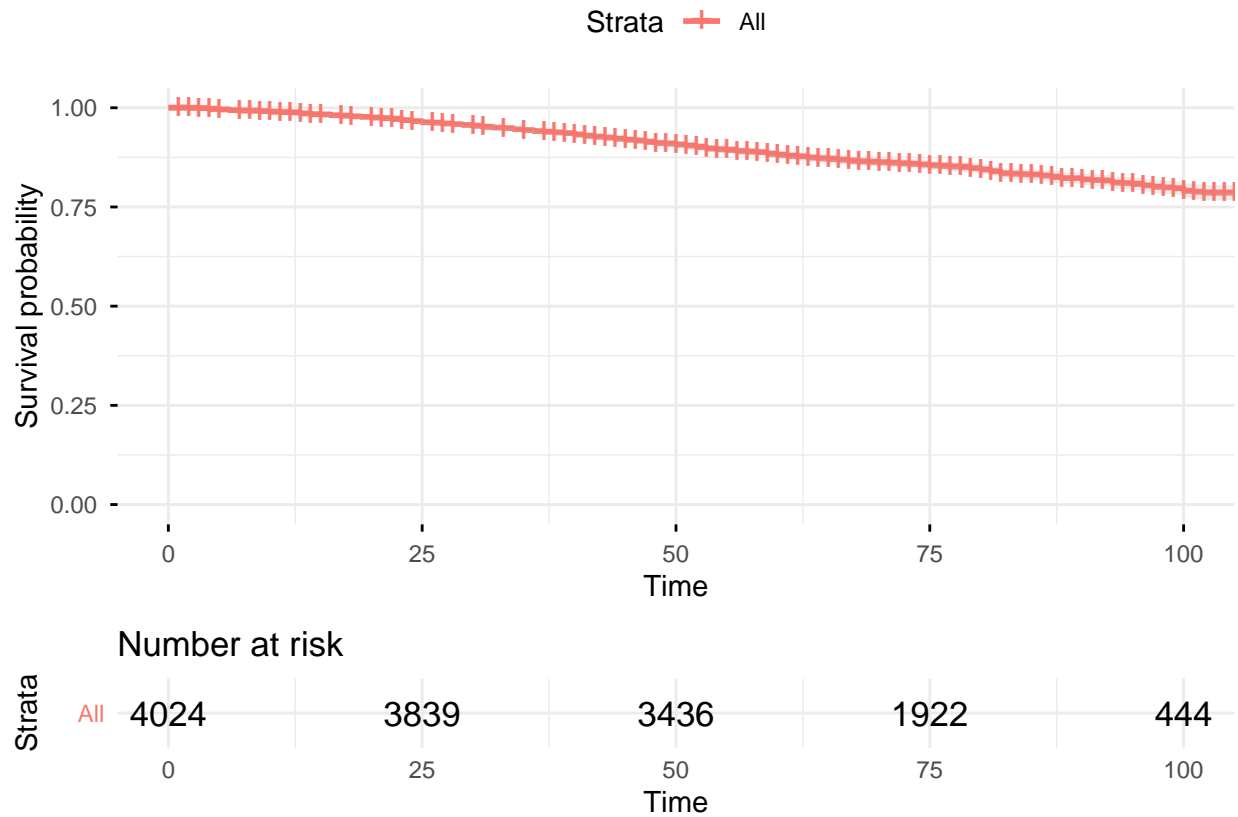
```
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * colour : "Strata"
```
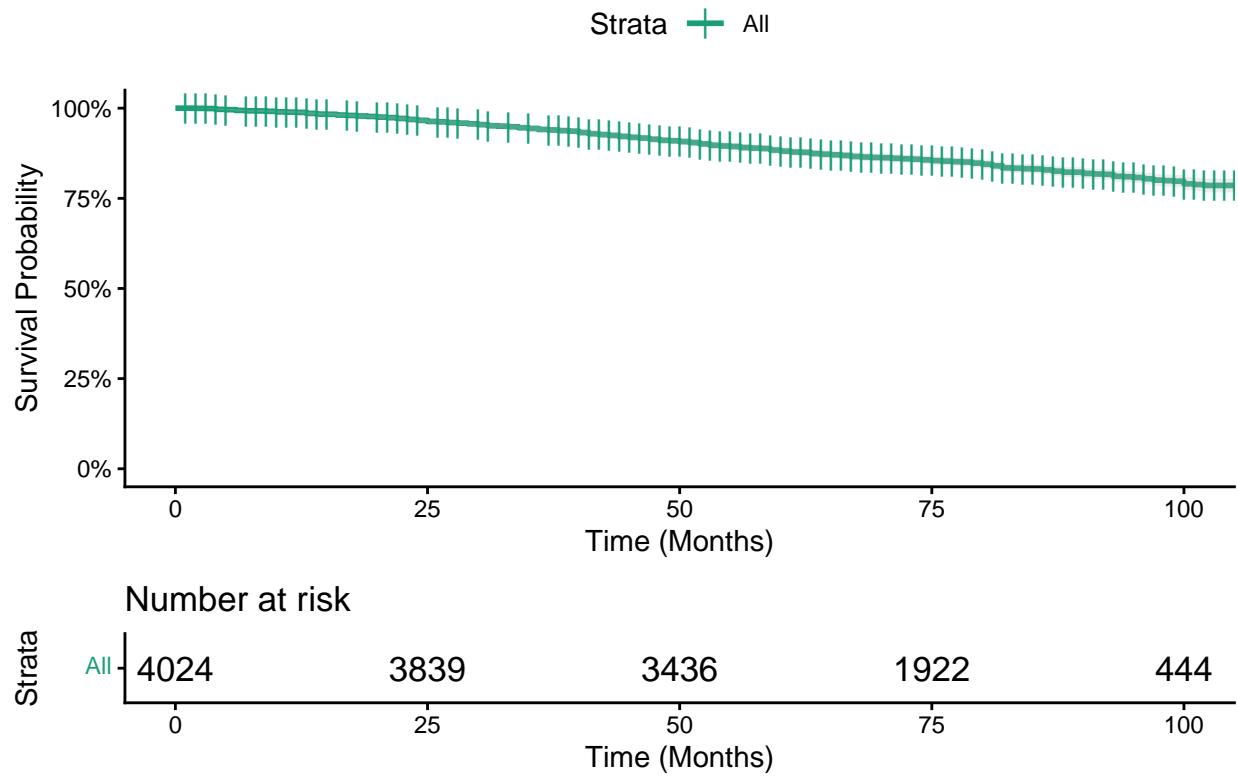

Additional way to plot survival curve

```
ggsurvplot(
  fit,
  data = data,
  conf.int = TRUE,
  risk.table = TRUE,
  surv.scale = "percent",
  xlab = "Time (Months)",
  ylab = "Survival Probability",
  title = "Kaplan-Meier Survival Curve",
  palette = "Dark2",
  censor.shape = 124,
  censor.size = 4,
  ggtheme = theme_classic()
)
```

```
## Ignoring unknown labels:
## * fill : "Strata"
```

```
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * fill : "Strata"
## Ignoring unknown labels:
## * colour : "Strata"
```

Kaplan−Meier Survival Curve

Strata ── All

	0	25	50	75	100

Survival Probability

Number at risk

Strata

All   4024   3839   3436   1922   444

Time (Months)

Characteristic	N = 4,024[1]
Age	54 (47, 61)
Race	
Black	291 (7.2%)
Other	320 (8.0%)
White	3,413 (85%)
Marital.Status	
Divorced	486 (12%)
Married	2,643 (66%)
Separated	45 (1.1%)
Single	615 (15%)
Widowed	235 (5.8%)
X	0 (NA%)
Unknown	4,024
T.Stage	
T1	1,603 (40%)
T2	1,786 (44%)
T3	533 (13%)
T4	102 (2.5%)
N.Stage	
N1	2,732 (68%)
N2	820 (20%)
N3	472 (12%)
X6th.Stage	
IIA	1,305 (32%)
IIB	1,130 (28%)
IIIA	1,050 (26%)
IIIB	67 (1.7%)
IIIC	472 (12%)
Grade	
Moderately differentiated; Grade II	2,351 (58%)
Poorly differentiated; Grade III	1,111 (28%)
Undifferentiated; anaplastic; Grade IV	19 (0.5%)
Well differentiated; Grade I	543 (13%)
A.Stage	
Distant	92 (2.3%)
Regional	3,932 (98%)
Tumor.Size	25 (16, 38)
Estrogen.Status	
Negative	269 (6.7%)
Positive	3,755 (93%)
Progesterone.Status	
Negative	698 (17%)
Positive	3,326 (83%)
Regional.Node.Examined	14 (9, 19)
Reginol.Node.Positive	2 (1, 5)
Survival.Months	73 (56, 90)
Status	616 (15%)

[1] Median (Q1, Q3); n (%)