Homework 7

James Young

Please do the following problems from the text book R Handbook and stated.

Packages used: survival, ISWR, HSAUR3, coin, party, survminer, gridExtra

Collaborators: Alex Soupir

Resources used: StackOverflow

1. An investigator collected data on survival of patients with lung cancer at Mayo Clinic. The investigator would like you, the statistician, to answer the following questions and provide some graphs. Use the **cancer** data located in the **survival** package.

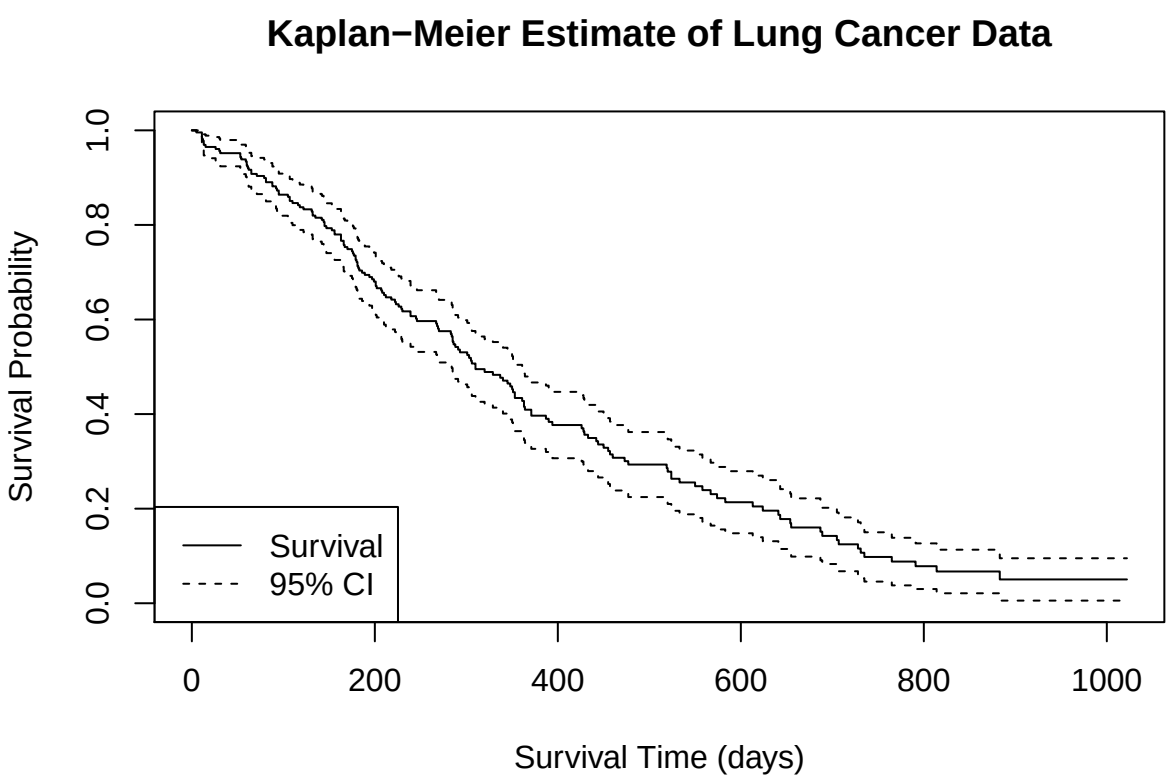
    a. What is the probability that someone will survive past 300 days?

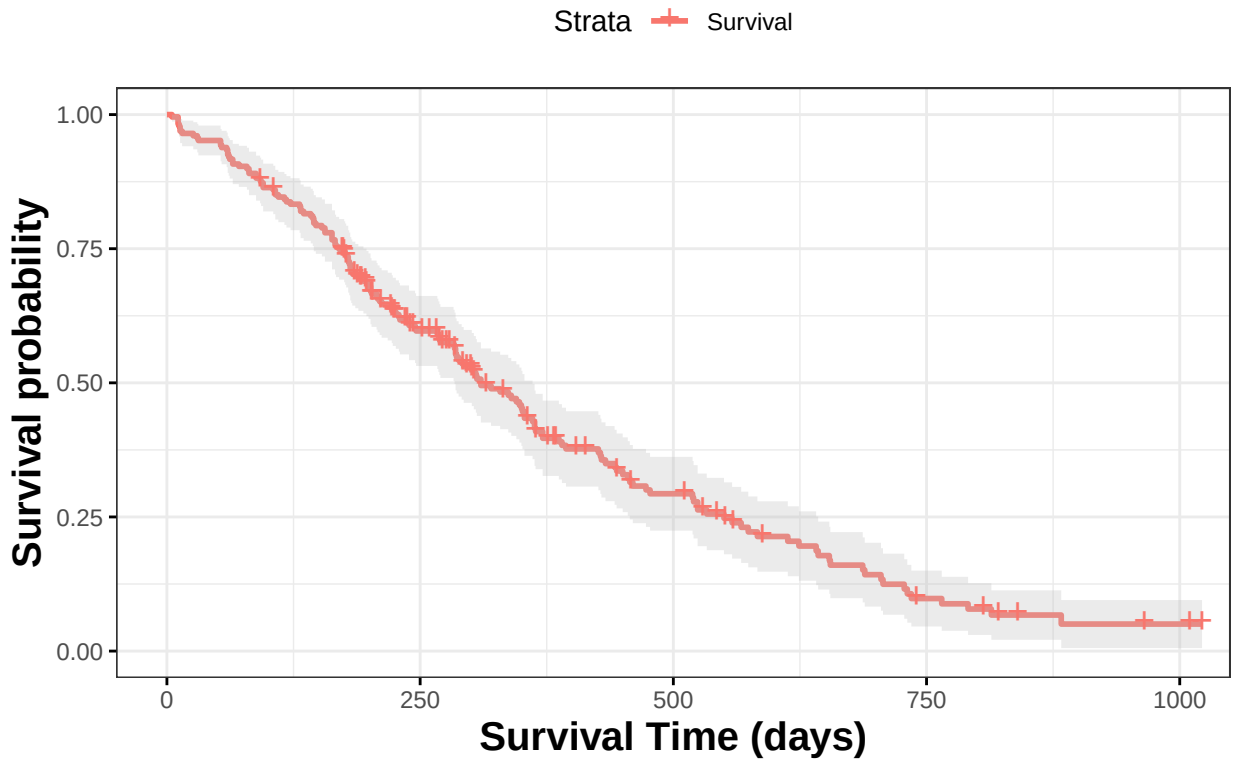
Patients who died received status of 2, so status in Surv was set to == 2. The probability of patient from this dataset living past 300 days is 0.531.

```
## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = cancer,
##     conf.int = 0.95, conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     92     101    0.531  0.0346        0.463        0.599
```

```
b. Provide a graph, including 95% confidence limits, of the Kaplan-Meier estimate of the entire study.
```

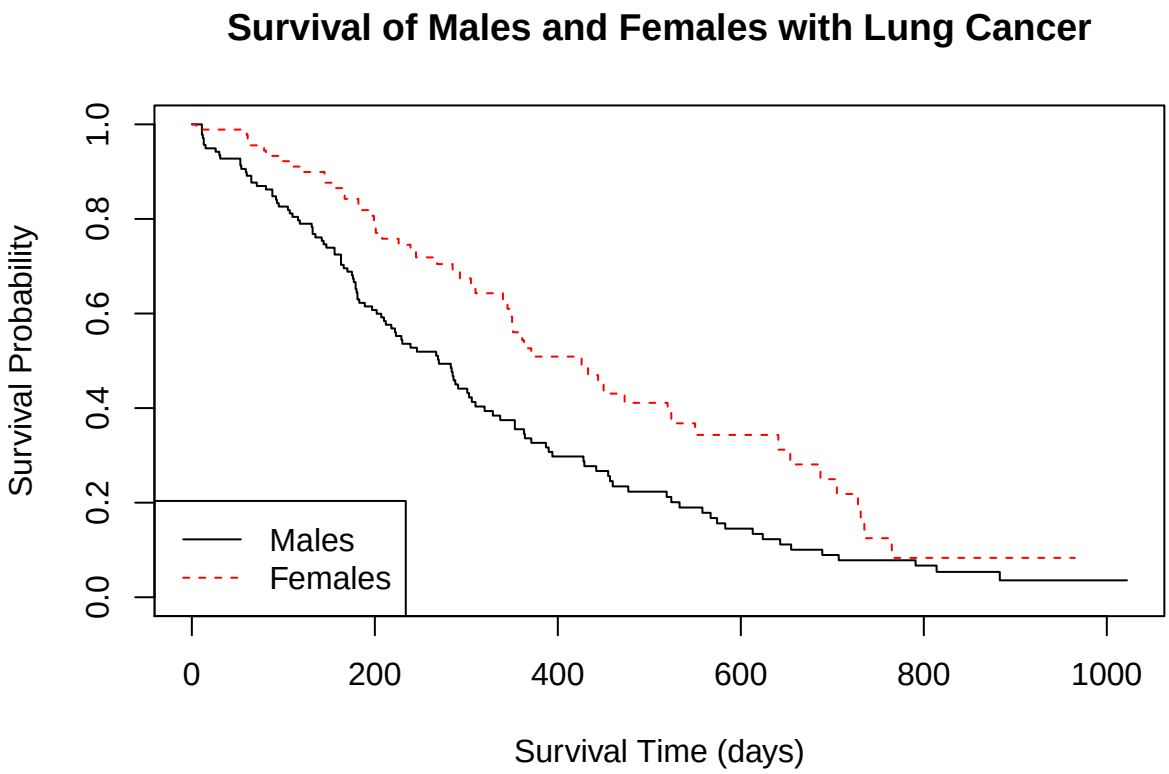
The requested graph is provided below in a base r version and ggplot version. The ggplot version is more aesthetically pleasing and easily customizable in my opinion.

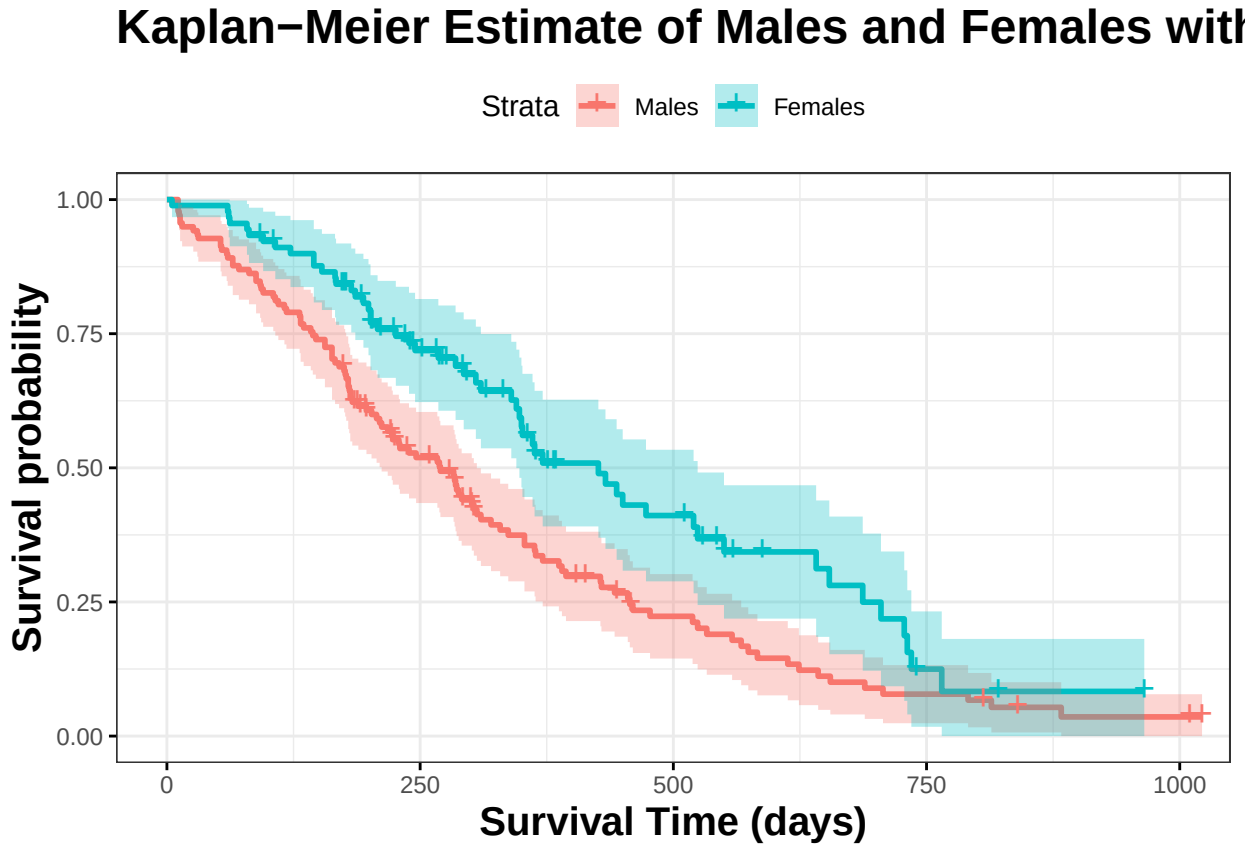

Kaplan−Meier Estimate of Lung Cancer Data
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
Survival
95% CI
0
200
400
600
800
1000
Survival Time (days)



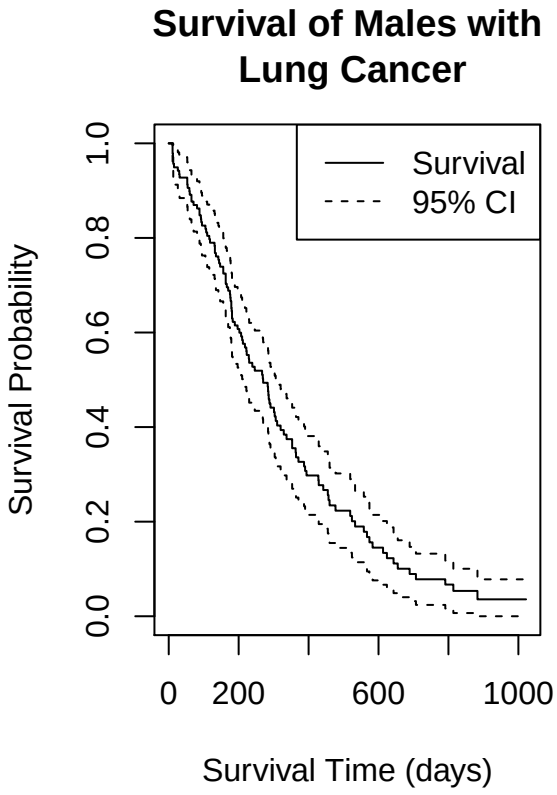
c. Is there a difference in the survival rates between males and females? Provide a formal statistical

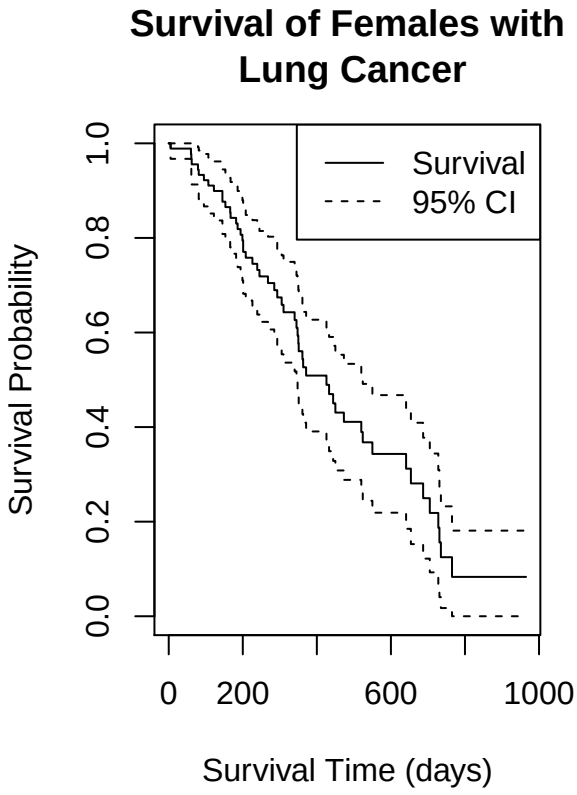
The first survival plot demonstrates females have a higher probability of survival until roughly 800 days. Males have a 0.4411 probability of surviving past 300 days and females have a probability of 0.674 of surviving past 300 days. Males and females had a probability of 0.0781 and 0.125 of surviving past 750 days, respectively. survdiff() was used to calculate a chisq of males vs females survival rate, which had a p-value of 0.001, showing a statistically significant difference in survival between males and females, with females having the good fortune in this case.

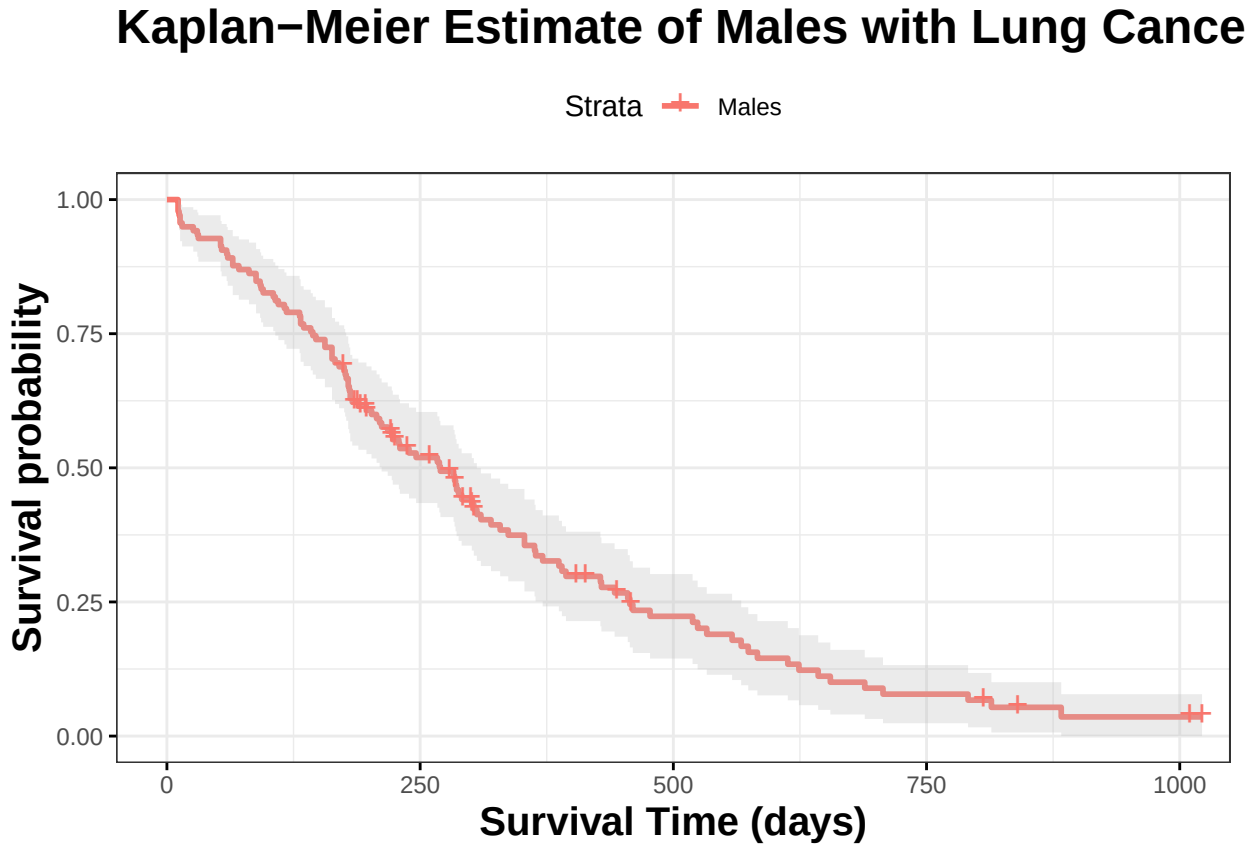

Survival of Males and Females with Lung Cancer
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
Males
Females
0
200
400
600
800
1000
Survival Time (days)

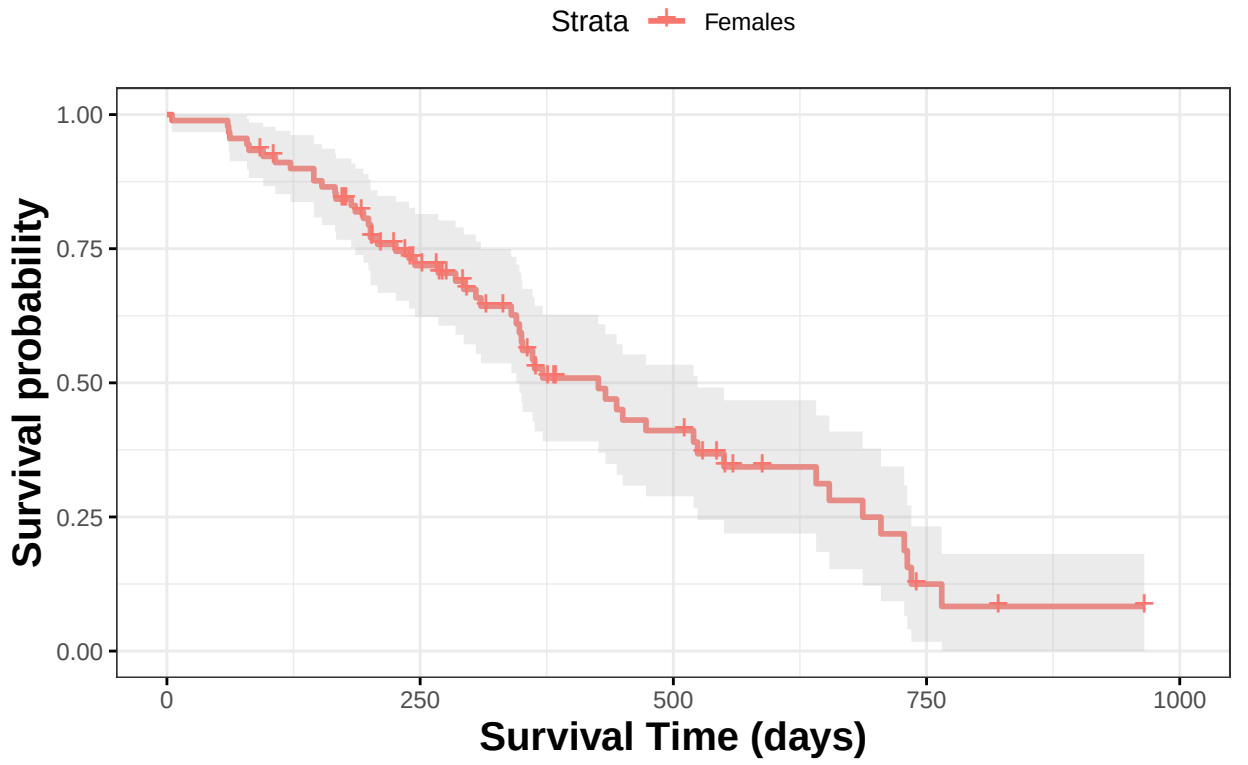



```
##
## Probability of a male living past 300 and 750 days:

## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = m2, conf.int = 0.95,
##     conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     49      74   0.4411  0.0439       0.3550        0.527
##   750      7      35   0.0781  0.0276       0.0239        0.132

##
## Probability of a female living past 300 and 750 days:

## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = fem, conf.int = 0.95,
##     conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     43      27    0.674  0.0523       0.5717        0.777
##   750      3      25    0.125  0.0549       0.0173        0.232
```


Survival of Males with Lung Cancer
Survival
95% CI
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
200
600
1000
Survival Time (days)


Survival of Females with Lung Cancer
Survival
95% CI
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
200
600
1000
Survival Time (days)


Kaplan−Meier Estimate of Males with Lung Cance
Strata
Males
1.00
0.75
0.50
0.25
0.00
Survival probability
0
250
500
750
1000
Survival Time (days)



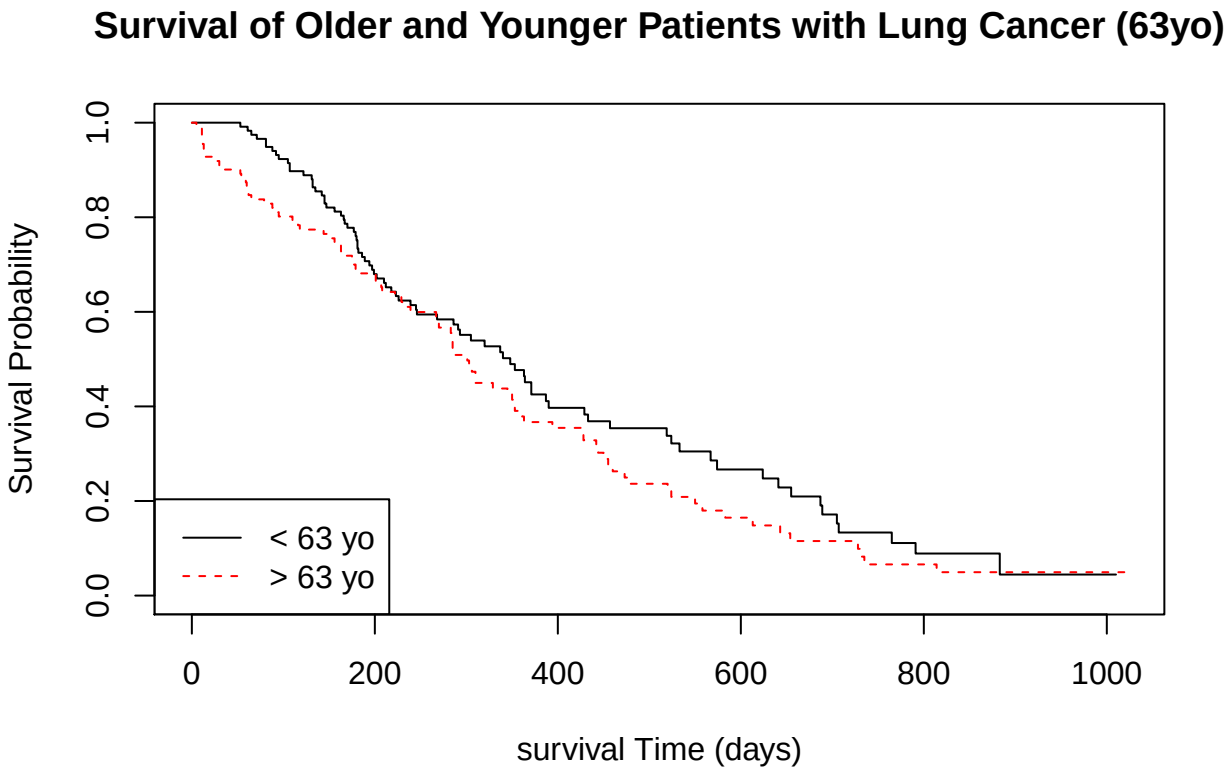

```
##
## Chisq test between males and females:

## Call:
## survdiff(formula = Surv(time, status == 2) ~ sex, data = cancer)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1 138      112     91.6      4.55      10.3
## sex=2  90       53     73.4      5.68      10.3
##
##  Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

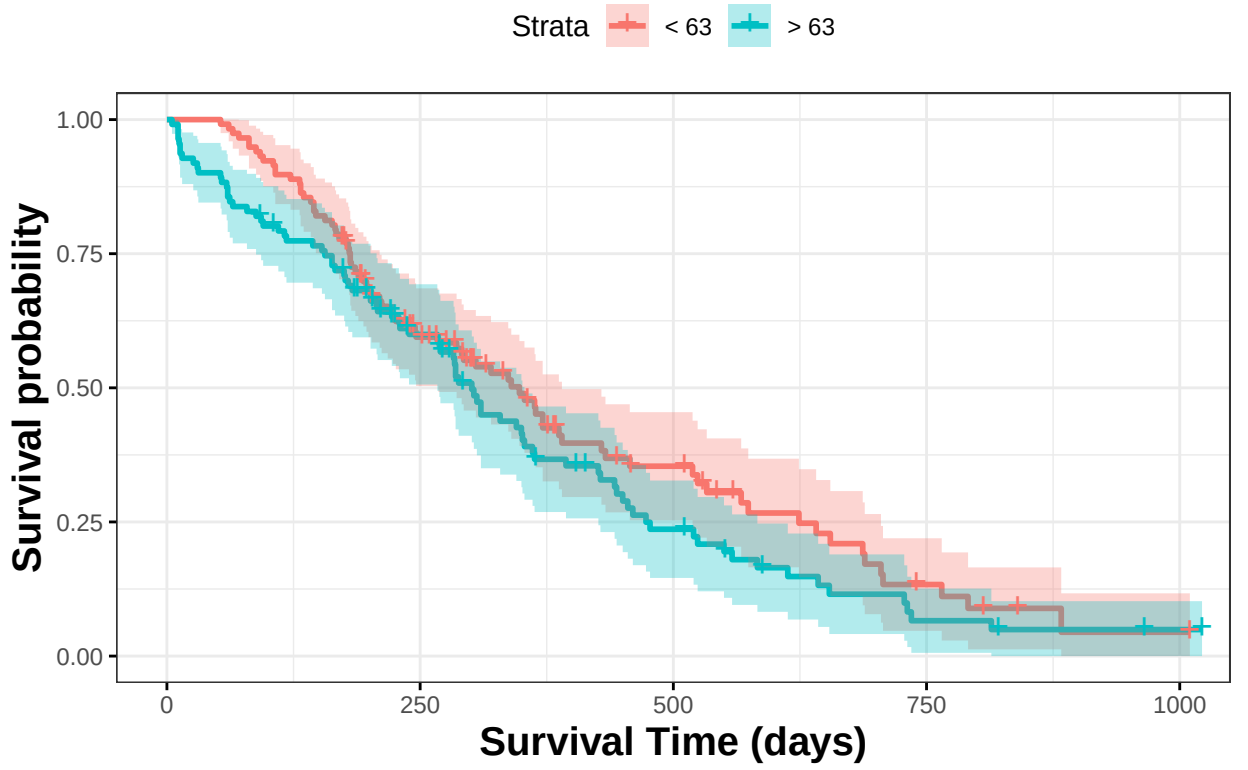
d. Is there a difference in the survival rates for the older half of the group versus the younger half?

The median age of patients was found to be 63 which served as the split point for comparing the survival rates of the older half verse the younger half of patients. The first plot appears to show relatively less difference in survival due to age then the difference we saw due to gender. The survival rates for the 2 groups was found with survfit() at times of 300 and 750 days. The younger group has a 0.551 probability of surviving past 300 days while the older group has a probability of 0.509 of surviving past 300 days. At 750 days, the older group has a survival probability of 0.066 and the younger group has a survival probability of 0.133. survdiff() was used to compute a chi-squared p-value of 0.2, meaning there is not a statistically significant difference in survival rates between the older and younger half of lung cancer patients in this data.

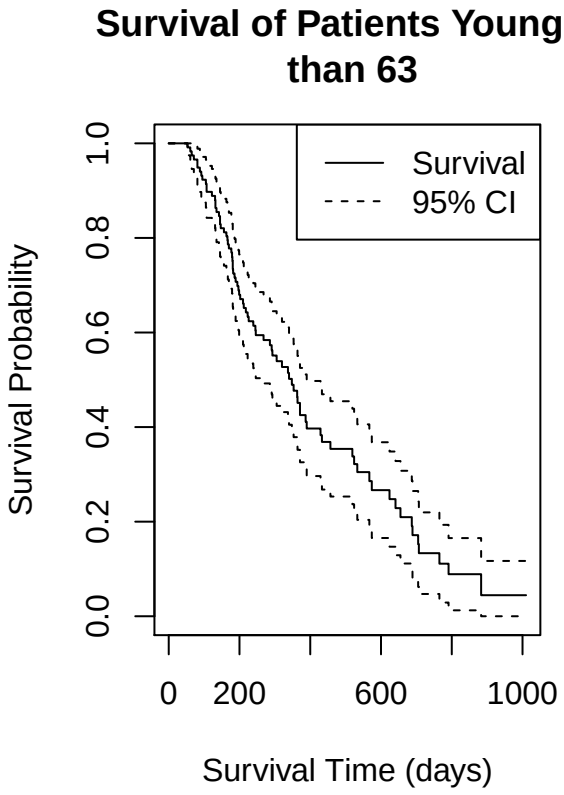
```
## Finding median age of cancer patients:
```

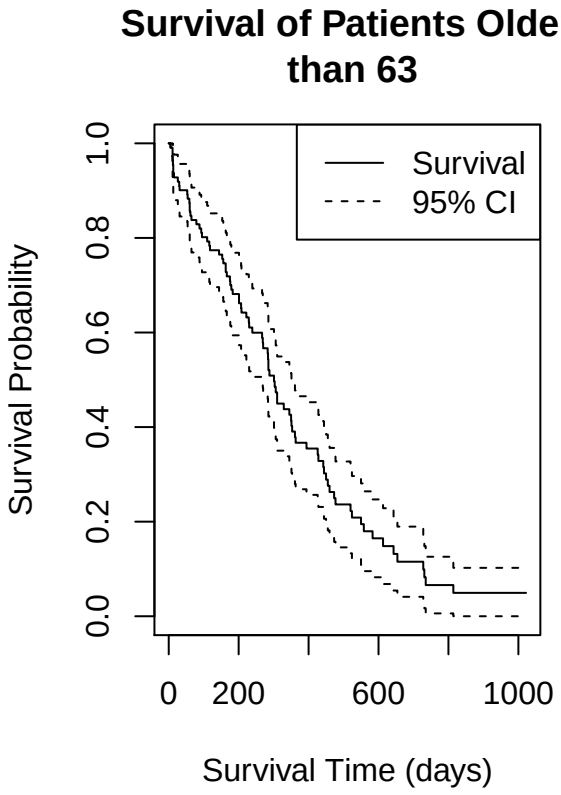
```
## [1] 63
```

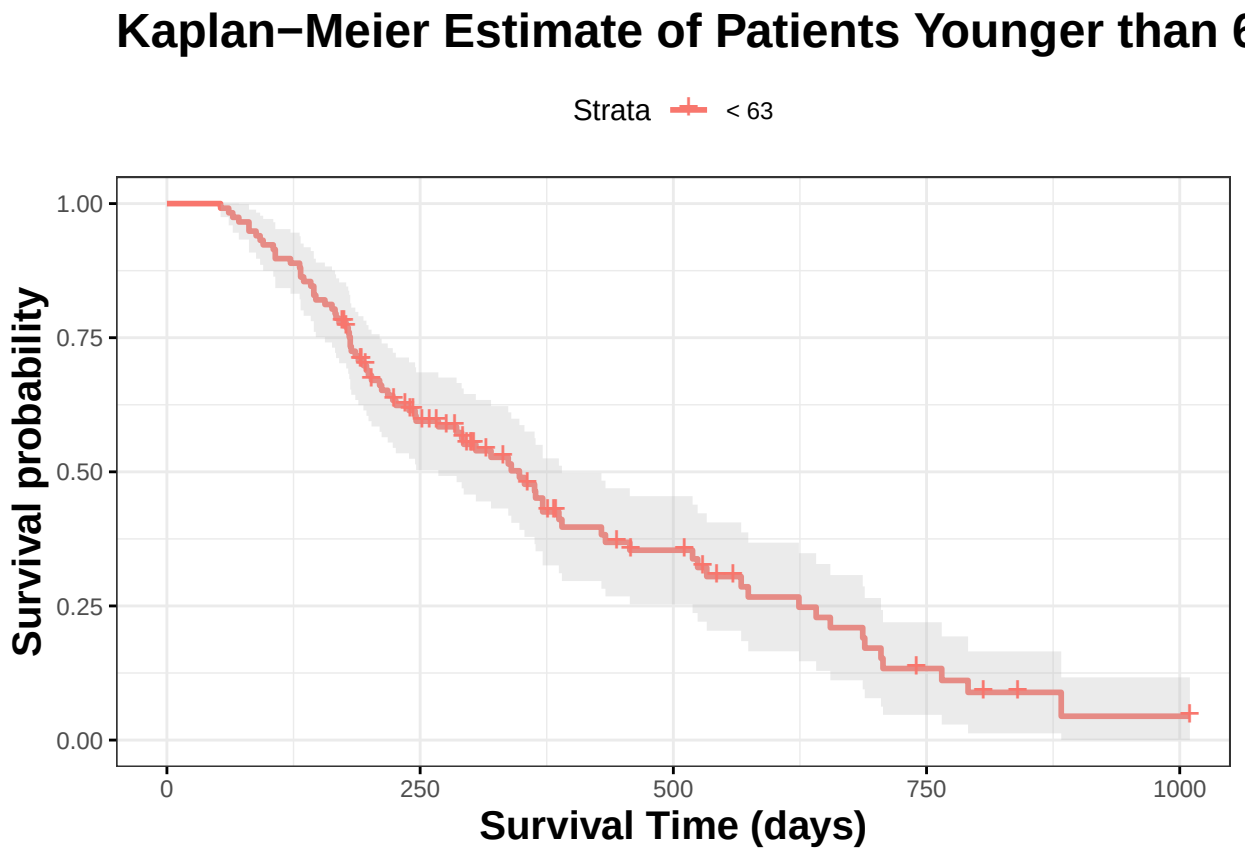


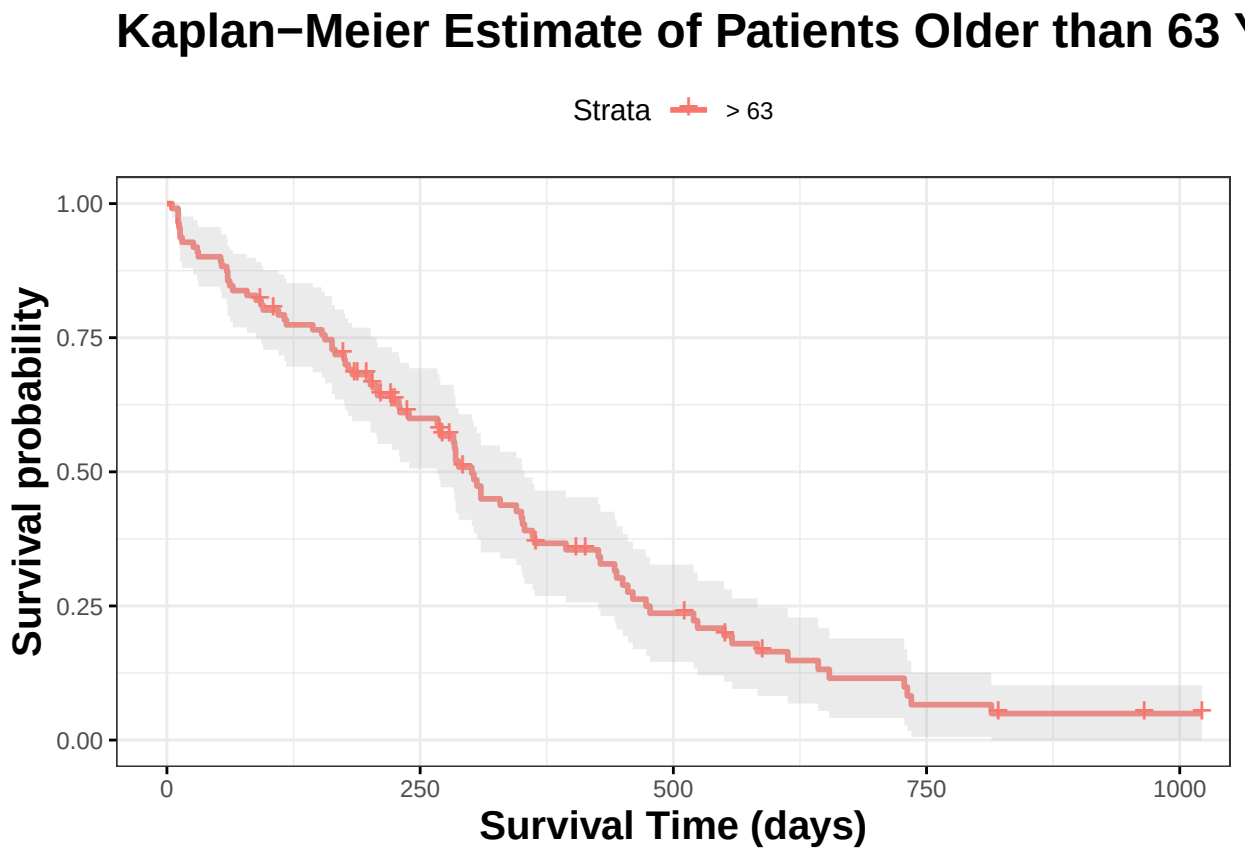



```
##
## Probability of younger than 63 living past 300 and 750 days:

## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = lm3, conf.int = 0.95,
##     conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     49      50    0.551  0.0478       0.4576        0.645
##   750      6      27    0.133  0.0440       0.0471        0.220

##
## Probability of those older than 63 living past 300 and 750 days:

## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = gm3, conf.int = 0.95,
##     conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     43      51   0.5089  0.0501      0.41060        0.607
##   750      4      33   0.0659  0.0306      0.00601        0.126
```


Survival of Patients Younger than 63
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
Survival
95% CI
0
200
600
1000
Survival Time (days)


Survival of Patients Older than 63
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
Survival
95% CI
0
200
600
1000
Survival Time (days)


Kaplan−Meier Estimate of Patients Younger than 6
Strata
< 63
Survival probability
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
Survival Time (days)



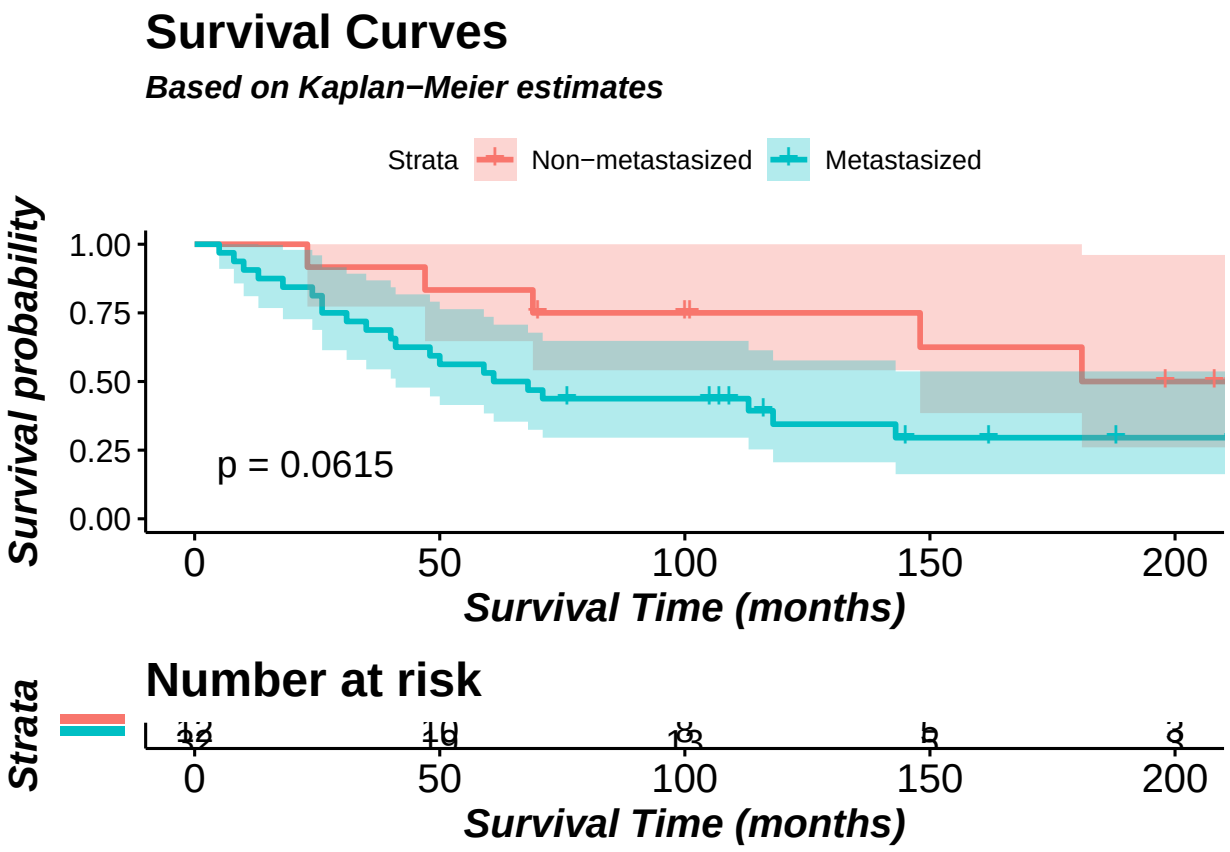

```
##
## Chisq test between younger and older:

## Call:
## survdiff(formula = Surv(time, status == 2) ~ agecat, data = cancer)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## agecat=<63 117       80     88.8     0.865      1.88
## agecat=>63 111       85     76.2     1.007      1.88
##
##  Chisq= 1.9  on 1 degrees of freedom, p= 0.2
```

Adding a little spacing to fulfill requirement of having question 2 on a single page.

Adding a little spacing to fulfill requirement of having question 2 on a single page.
Adding a little spacing to fulfill requirement of having question 2 on a single page.
Adding a little spacing to fulfill requirement of having question 2 on a single page.
Adding a little spacing to fulfill requirement of having question 2 on a single page.
Adding a little spacing to fulfill requirement of having question 2 on a single page.
Adding a little spacing to fulfill requirement of having question 2 on a single page.
Adding a little spacing to fulfill requirement of having question 2 on a single page.
Adding a little spacing to fulfill requirement of having question 2 on a single page.
Adding a little spacing to fulfill requirement of having question 2 on a single page.
Adding a little spacing to fulfill requirement of having question 2 on a single page.

2. A healthcare group has asked you to analyse the **mastectomy** data from the **HSAUR3** package, which is the survival times (in months) after a mastectomy of women with breast cancer. The cancers are classified as having metastasized or not based on a histochemical marker. The healthcare group requests that your report should not be longer than one page, and must only consist of one plot, one table, and one paragraph. Do the following:

   a. Plot the survivor functions of each group only using GGPlot, estimated using the Kaplan-Meier estimate.



b. Use a log-rank test to compare the survival experience of each group more formally. Only present a f

```
## Log-Rank Test Results:

##                                      Formula Z.value P.value
## 1 Surv(time, event == TRUE) by metastasized  1.8667 0.06146
```

c. Write one paragraph summarizing your findings and conclusions.

The patients identified as having metastasized cancer appear to be at a higher risk of death than patients identified as non-metastisized based on the Survival plot. The non-metastized patients have a much higher upper-bound on their confidence interval across the 50 months. However, there is overlap between the confidence intervals of metastisized and non-metastisized patients. To give more robust insight, a log-rank test was completed which showed that the difference between the two groups is not statistically significant at the 0.05 level, although it is very close (p-value ~ 0.06).