Code ▾

Breast Cancer Data Preparation and Visualization

dataset source: https://archive.ics.uci.edu/dataset/14/breast+cancer (https://archive.ics.uci.edu/dataset/14/breast+cancer)

Hide

```
library(tidyverse)
```

1. Data Preprocessing

Hide

```
# read the file as a dataframe
breast.data = read.csv("breast-cancer.data", header = FALSE)
```

Hide

```
# View the first and last six rows
head(breast.data, n=6)
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	▶
	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>	<chr>	<chr>	
1	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	
2	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	
3	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	
4	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	
5	no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	
6	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	left	left_low	

6 rows | 1-10 of 10 columns

Hide

```
tail(breast.data, n=6)
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	▶
	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<int>	<chr>	<chr>	
281	recurrence-events	50-59	ge40	40-44	6-8	yes	3	left	left_low	
282	recurrence-events	30-39	premeno	30-34	0-2	no	2	left	left_up	
283	recurrence-events	30-39	premeno	20-24	0-2	no	3	left	left_up	

	V1 <chr>	V2 <chr>	V3 <chr>	V4 <chr>	V5 <chr>	V6 <chr>	V7 <int>	V8 <chr>	V9 <chr>
284	recurrence-events	60-69	ge40	20-24	0-2	no	1	right	left_up
285	recurrence-events	40-49	ge40	30-34	3-5	no	3	left	left_low
286	recurrence-events	50-59	ge40	30-34	3-5	no	3	left	left_low

6 rows | 1-10 of 10 columns

Hide

```
# rename the columns
colnames(breast.data) = c("class", "age", "menopause", "tumor.size",
                          "inv.nodes", "node.caps","deg.malignancy",
                          "breast","breast.quad","irradiation")
head(breast.data)
```

	class <chr>	age <chr>	menopa... <chr>	tumor.size <chr>	inv.nodes <chr>	node.caps <chr>	deg.malignancy <int>
1	no-recurrence-events	30-39	premeno	30-34	0-2	no	3
2	no-recurrence-events	40-49	premeno	20-24	0-2	no	2
3	no-recurrence-events	40-49	premeno	20-24	0-2	no	2
4	no-recurrence-events	60-69	ge40	15-19	0-2	no	2
5	no-recurrence-events	40-49	premeno	0-4	0-2	no	2
6	no-recurrence-events	60-69	ge40	15-19	0-2	no	2

6 rows | 1-9 of 10 columns

Hide

```
# check for missing values which are denoted by "?" in node-caps and breast-quad
head(breast.data[c(breast.data$node.caps == "?"|breast.data$breast.quad =="?"),])
```

	class <chr>	age <chr>	menopa... <chr>	tumor.size <chr>	inv.nodes <chr>	node.caps <chr>	deg.malignancy <int>
146	no-recurrence-events	40-49	premeno	25-29	0-2	?	2

	class <chr>	age <chr>	menopa… <chr>	tumor.size <chr>	inv.nodes <chr>	node.caps <chr>	deg.malignancy <int>
164	no-recurrence-events	60-69	ge40	25-29	3-5	?	1
165	no-recurrence-events	60-69	ge40	25-29	3-5	?	1
184	no-recurrence-events	50-59	ge40	30-34	9-11	?	3
185	no-recurrence-events	50-59	ge40	30-34	9-11	?	3
207	recurrence-events	50-59	ge40	30-34	0-2	no	3

Hide

```
# check the count of observations with missing values
nrow(breast.data[c(breast.data$node.caps == "?"|breast.data$breast.quad == "?"),])
```

```
[1] 9
```

Hide

```
# replace "?" to NA
new_breast_data = replace(breast.data,breast.data == "?",NA)
sum(is.na(new_breast_data))
```

```
[1] 9
```

Hide

```
# remove rows with missing values
new_breast_data = na.omit(new_breast_data)
```

Hide

```
# save it as a new csv file
write.csv(new_breast_data,"breast_cancer_new.csv")
```

2. Data Visualization

Hide

```
# factorize and relevel values of some variables
new_breast_data$class = factor(new_breast_data$class)
new_breast_data$menopause = factor(new_breast_data$menopause)
new_breast_data$node.caps = factor(new_breast_data$node.caps)
new_breast_data$breast = factor(new_breast_data$breast)
new_breast_data$breast.quad = factor(new_breast_data$breast.quad)
new_breast_data$irradiation = factor(new_breast_data$irradiation)
new_breast_data$age = factor(new_breast_data$age,
                         levels = c("20-29", "30-39" ,"40-49","50-59" ,"60-69",
                                    "70-79"))
new_breast_data$inv.nodes = factor(new_breast_data$inv.nodes,
                         levels = c("0-2", "3-5" ,"6-8","9-11" ,"12-14",
                                    "15-17","24-26"))
new_breast_data$deg.malignancy = factor(new_breast_data$deg.malignancy,
                                    levels = c("1","2","3"))
new_breast_data$tumor.size = factor(new_breast_data$tumor.size,
                                           levels = c("0-4","5-9", "10-14","15-19", "20-24","25-
29" ,"30-34" ,"35-39" ,"40-44" ,"45-49","50-54" ))
```

Hide

```
summary(new_breast_data)
```

```
                  class        age       menopause     tumor.size inv.nodes
 no-recurrence-events:196   20-29: 1   ge40   :123   30-34  :57   0-2  :209
 recurrence-events   : 81   30-39:36   lt40   :  5   25-29  :51   3-5  : 34
                            40-49:89   premeno:149   20-24  :48   6-8  : 17
                            50-59:91                 15-19  :29   9-11 :  7
                            60-69:55                 10-14  :28   12-14:  3
                            70-79: 5                 40-44  :22   15-17:  6
                                                     (Other):42   24-26:  1
 node.caps deg.malignancy   breast       breast.quad  irradiation
 no :221   1: 66         left :145   central  : 21   no :215
 yes: 56   2:129         right:132   left_low :106   yes: 62
           3: 82                     left_up  : 94
                                     right_low: 23
                                     right_up : 33
```

Hide

```
# create a variable to store column names of the dataset
variables = colnames(new_breast_data)
variables
```

```
 [1] "class"          "age"            "menopause"      "tumor.size"
 [5] "inv.nodes"      "node.caps"      "deg.malignancy" "breast"
 [9] "breast.quad"    "irradiation"
```

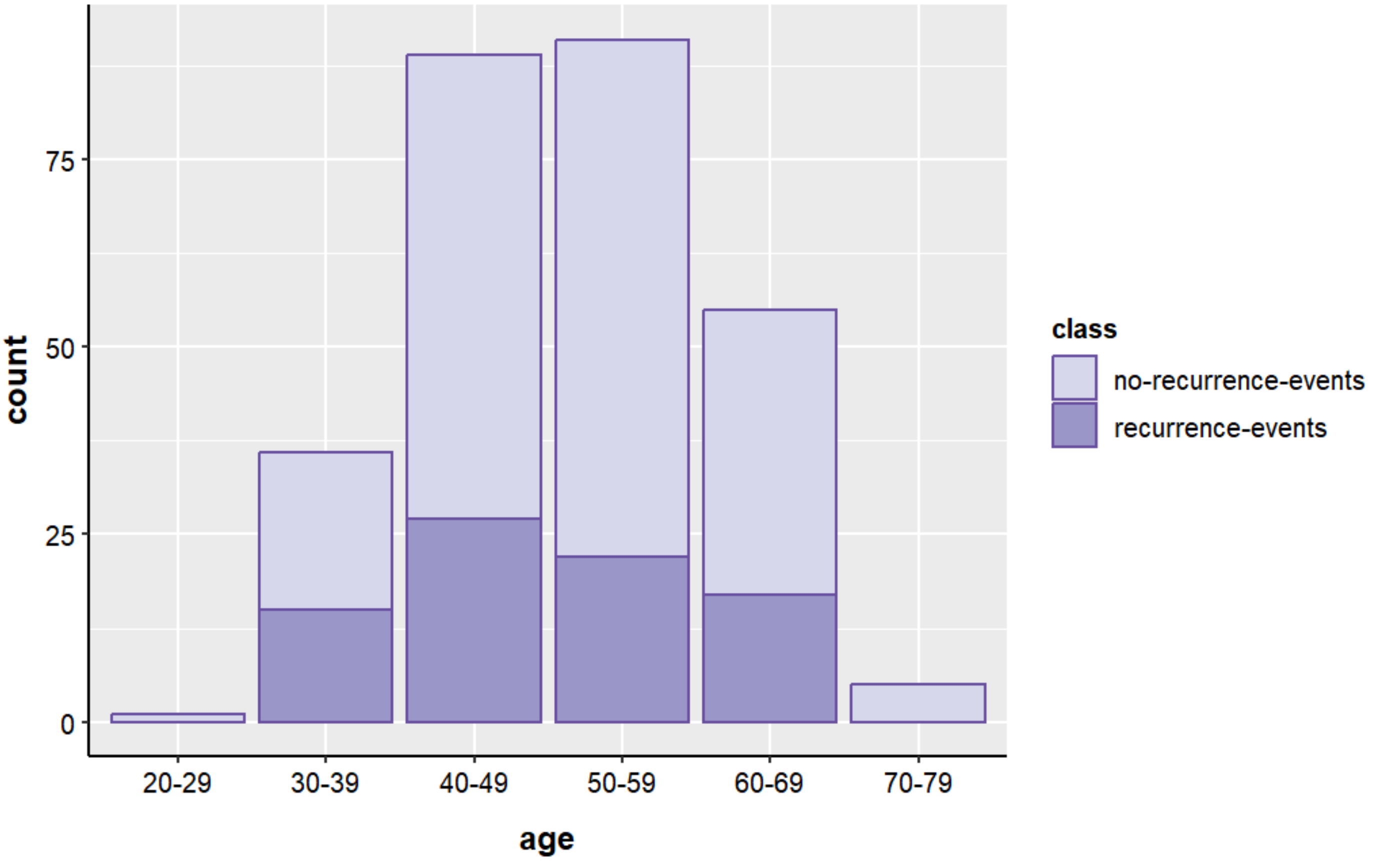
Hide

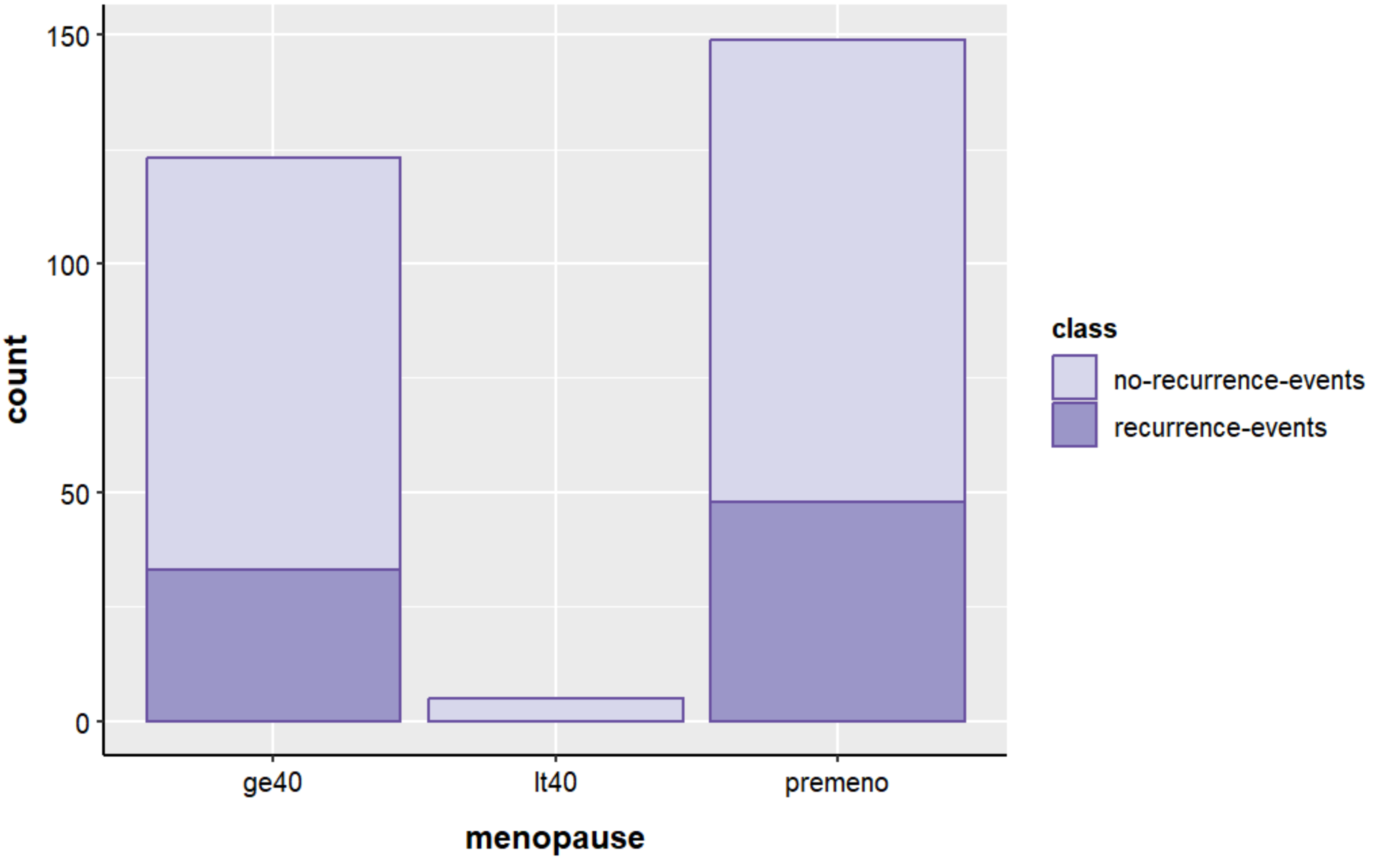
```
# create a function to make stacked barplot of breast cancer recurrence by each variable
stacked_barplot = function(dataset, x, fill, x_axis_name, legend_label)
  {imj = ggplot(dataset, aes(x = x, fill = fill)) +
          geom_bar(color = "#6A51A3") +
          scale_fill_manual(values = c("#DADAEB", "#9E9AC8"))+
          labs(x = x_axis_name, fill = legend_label) +
          theme(axis.line = element_line(colour="black",
                                         size=0.5, linetype="solid"),
              plot.title = element_text(face="bold", color="black",
                                        size=13, hjust=0.6, vjust=+1),
              axis.text.x = element_text(color="black", size=10),
              axis.text.y = element_text(color="black", size=10),
              axis.title.x = element_text(face="bold", color="black",
                                          size=12, ,hjust=0.5, vjust=-3,
                                          margin=margin(t=0,r=0,b=10,l=0)),
              axis.title.y = element_text(face="bold", color="black",
                                          size=12, vjust=+3, hjust=0.5,
                                          margin=margin(t=0,r=0,b=0,l=10)),
              legend.title = element_text(size=10,face='bold', color='black'),
              legend.text = element_text(size=10, color='black'))

  return (imj)
  }
```

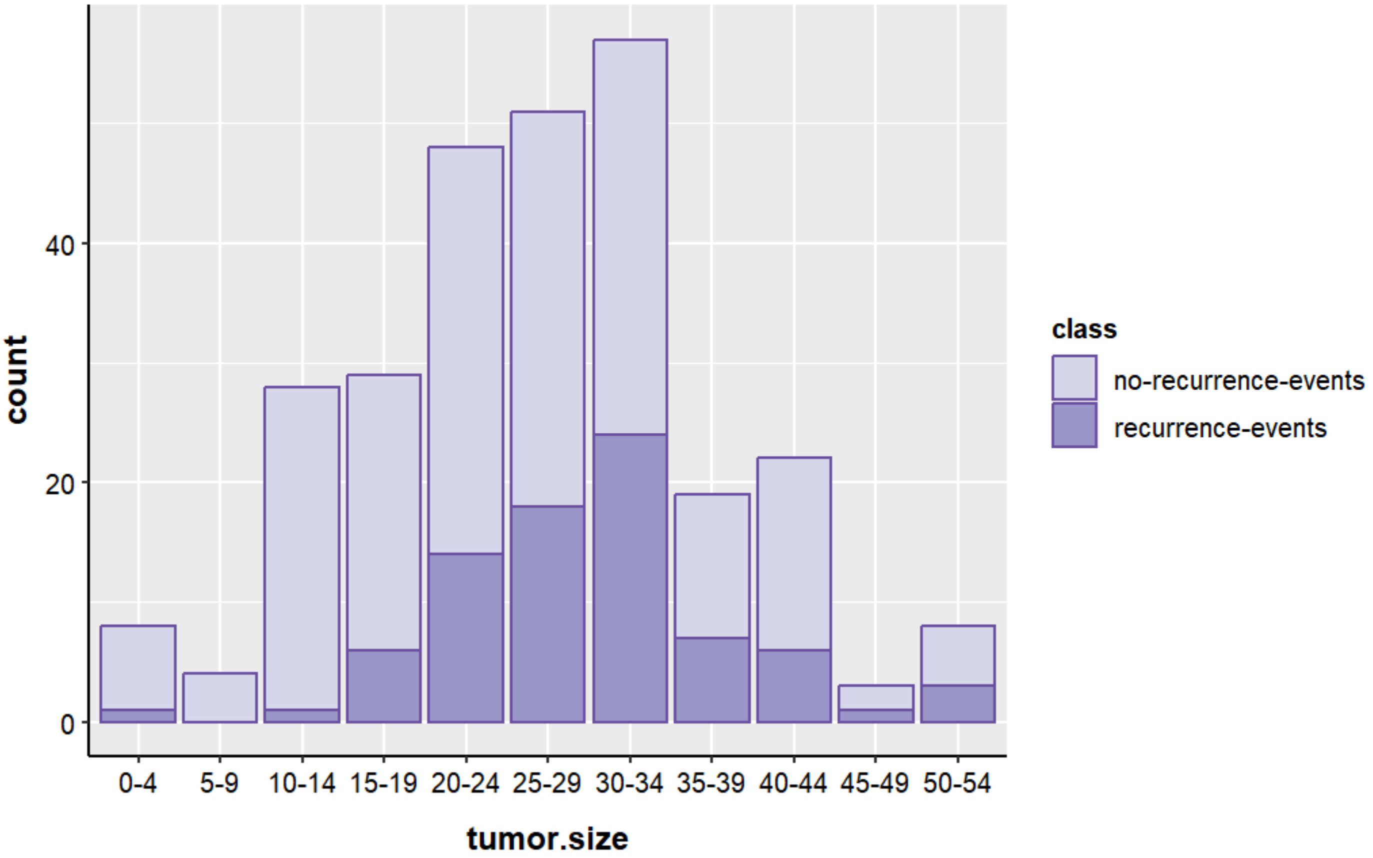
Hide

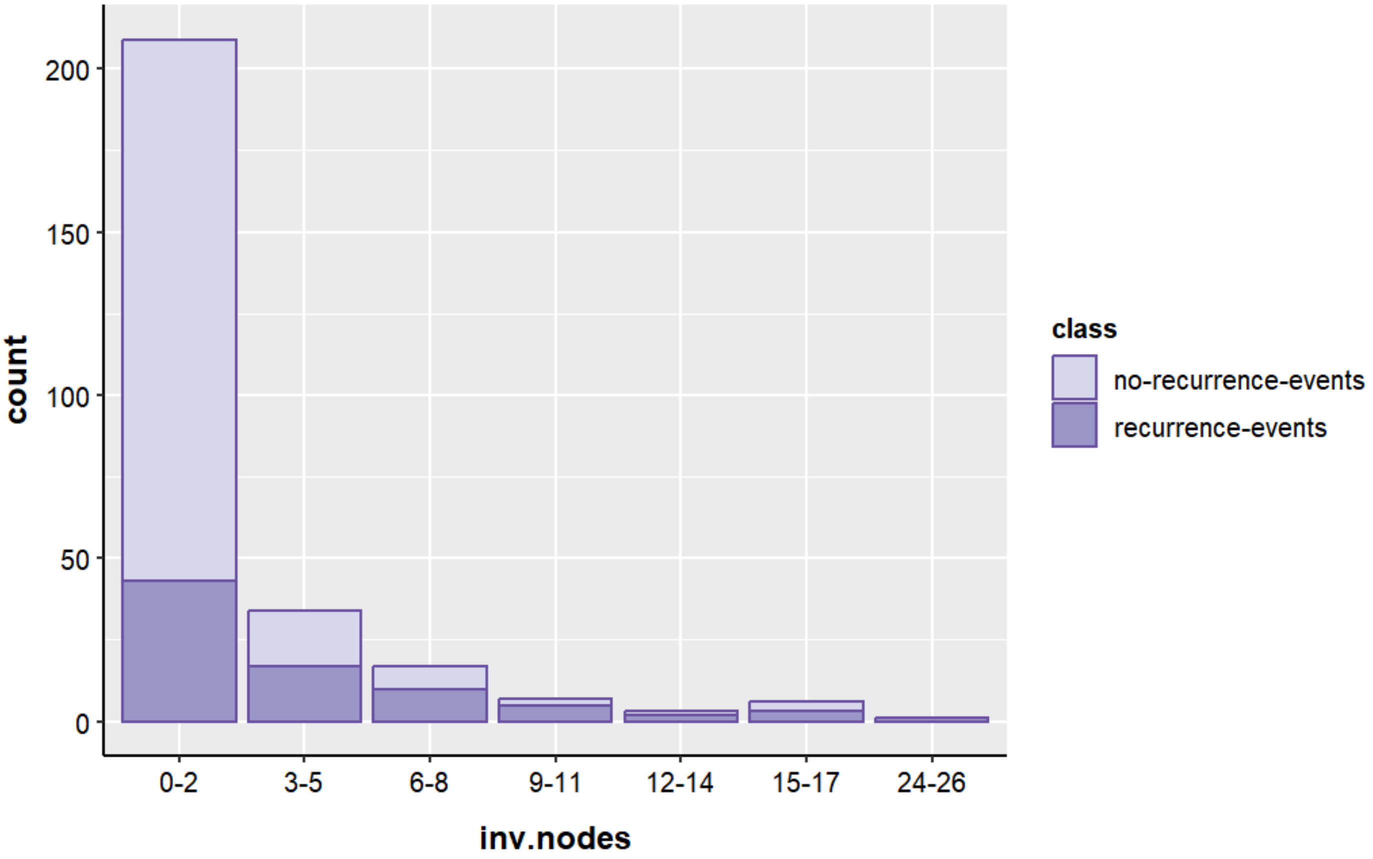
```
# create a stacked barplot for breast cancer recurrence by each predictor variable
for (i in 2:length(variables))
{
  imj = stacked_barplot(new_breast_data,new_breast_data[[i]],new_breast_data[[1]],
                        variables[i], variables[1])
  print(imj)
}
```

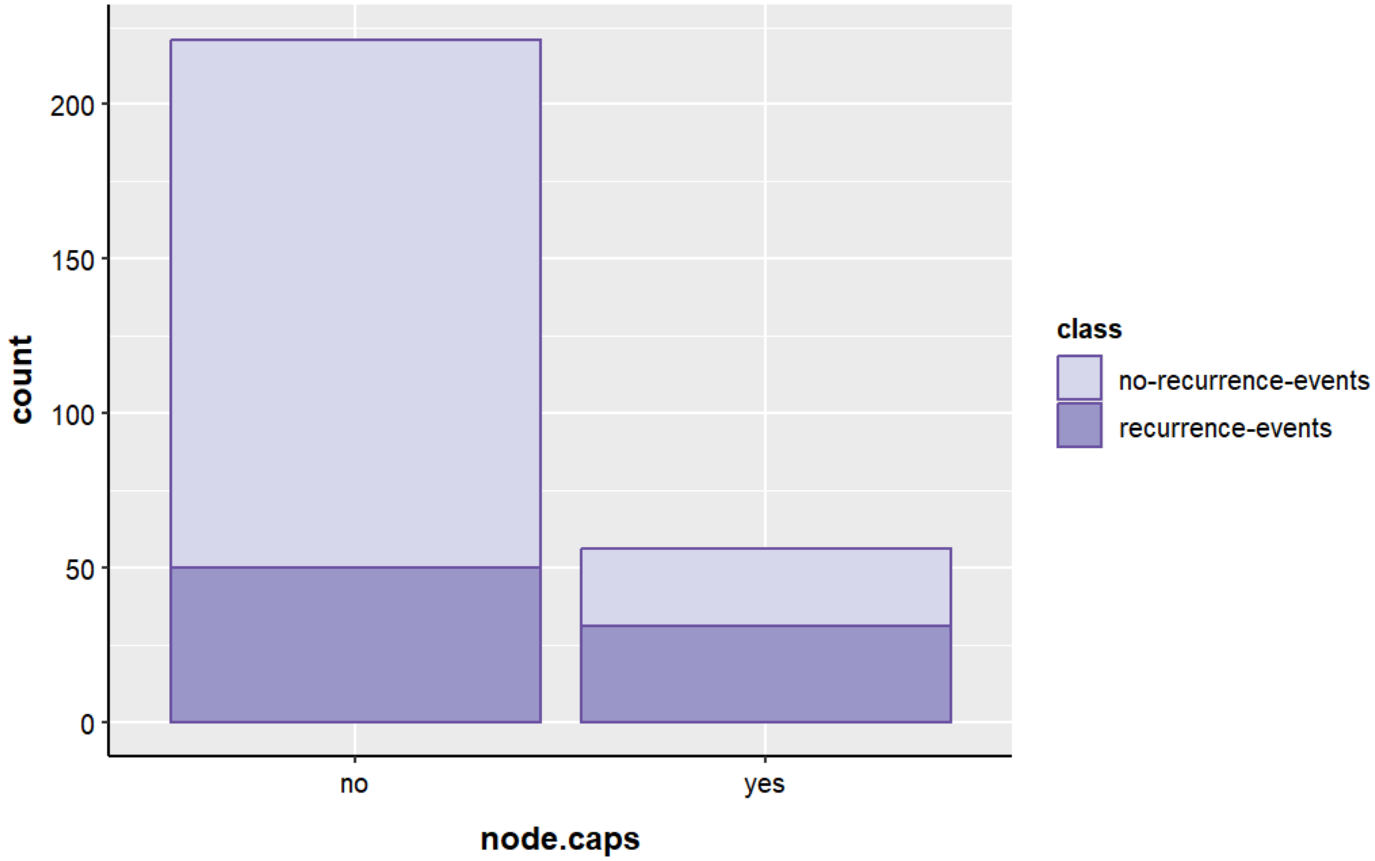
```
Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
Please use the `linewidth` argument instead.
```

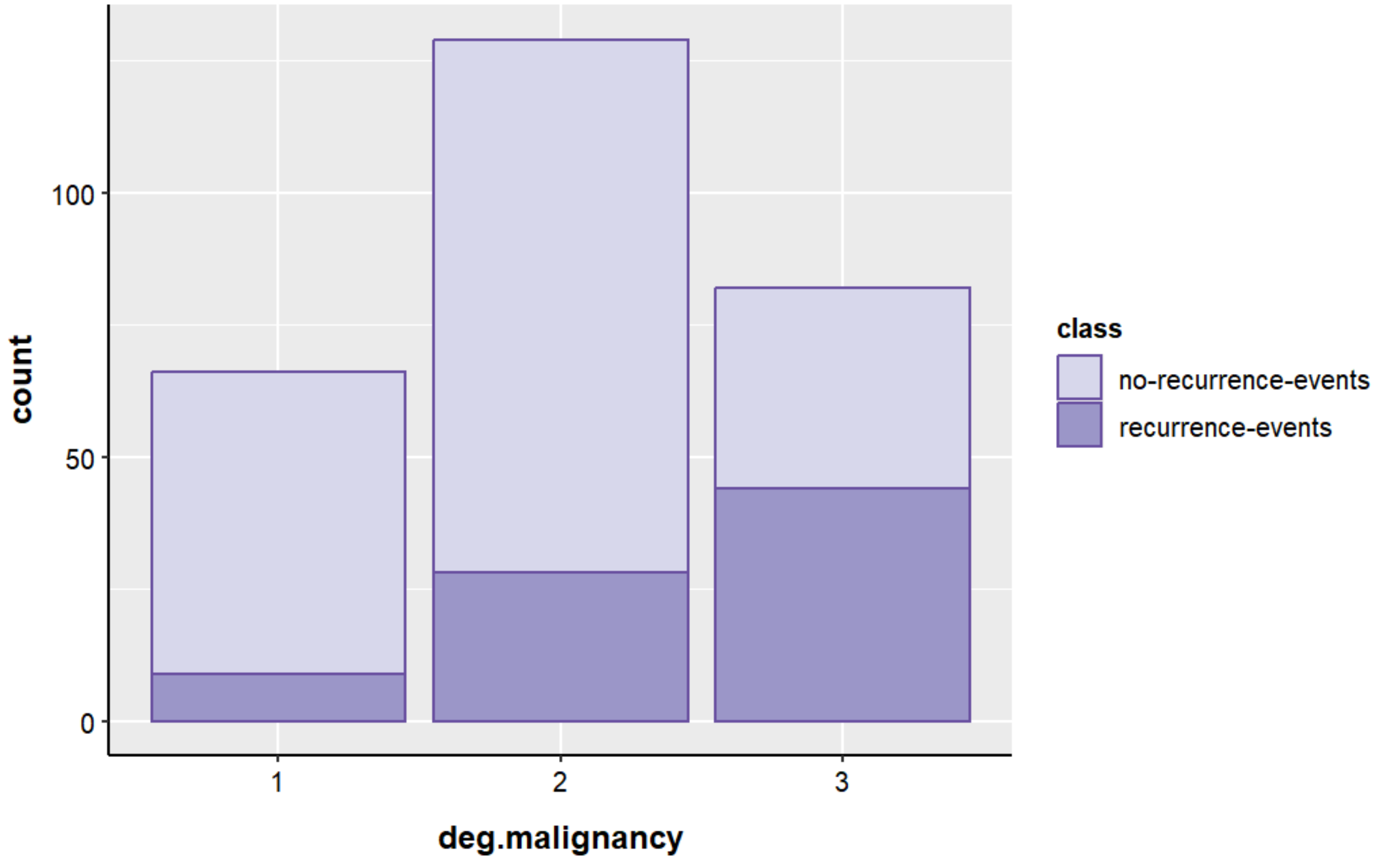

count
0
25
50
75
20-29
30-39
40-49
50-59
60-69
70-79
age
class
no-recurrence-events
recurrence-events

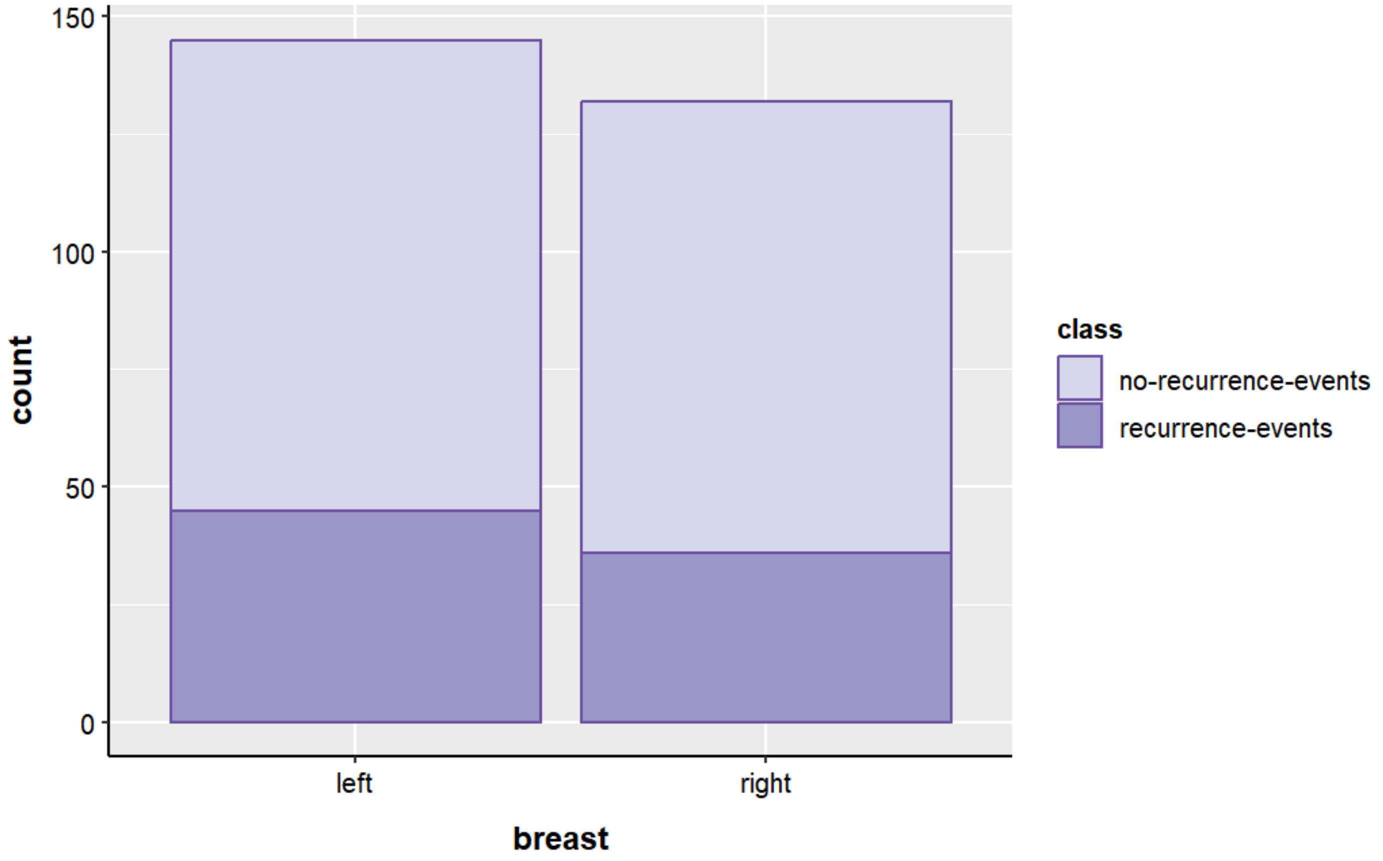

count
0
50
100
150
ge40
lt40
premeno
menopause
class
no-recurrence-events
recurrence-events

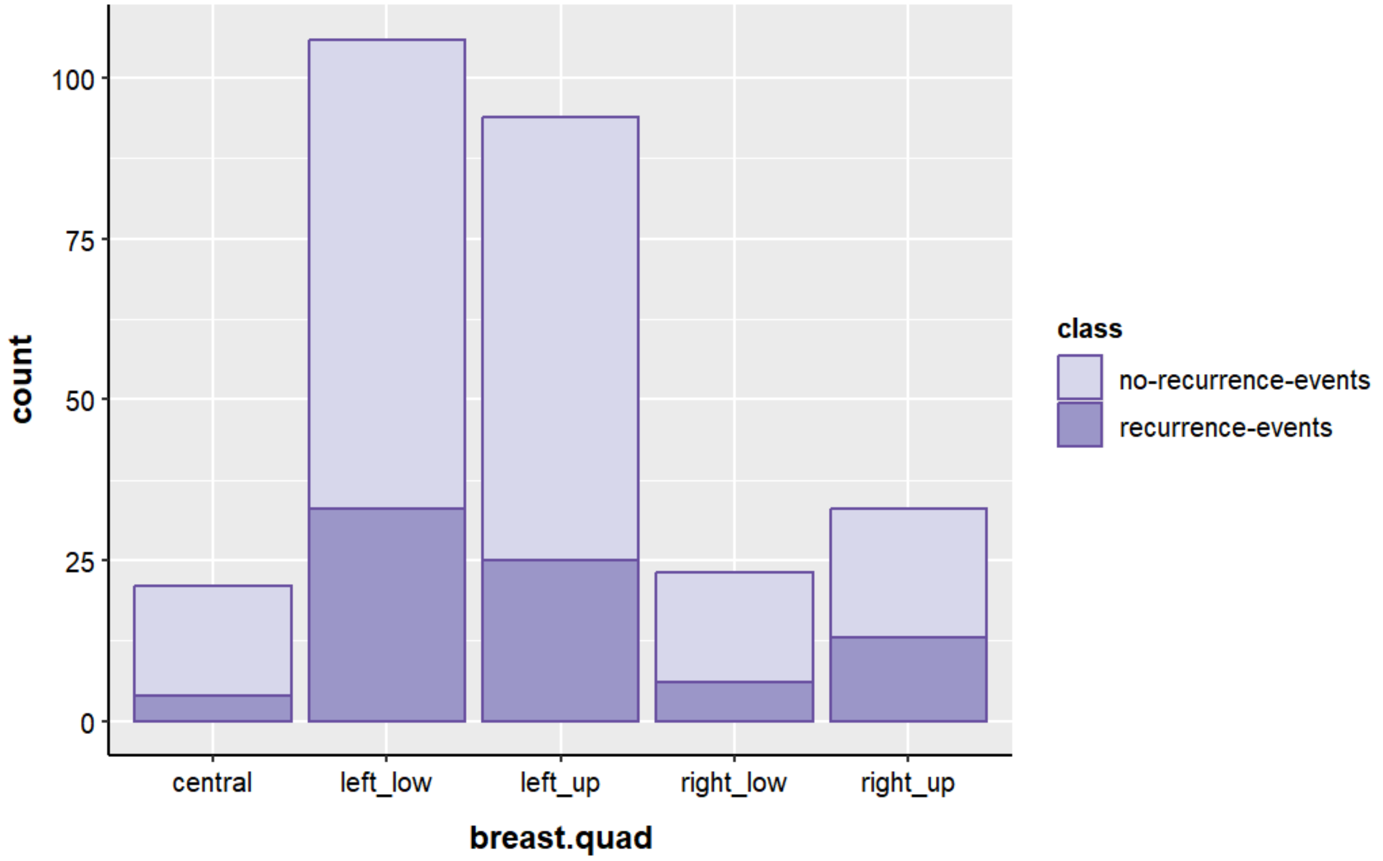

count
0
20
40
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
tumor.size
class
no-recurrence-events
recurrence-events

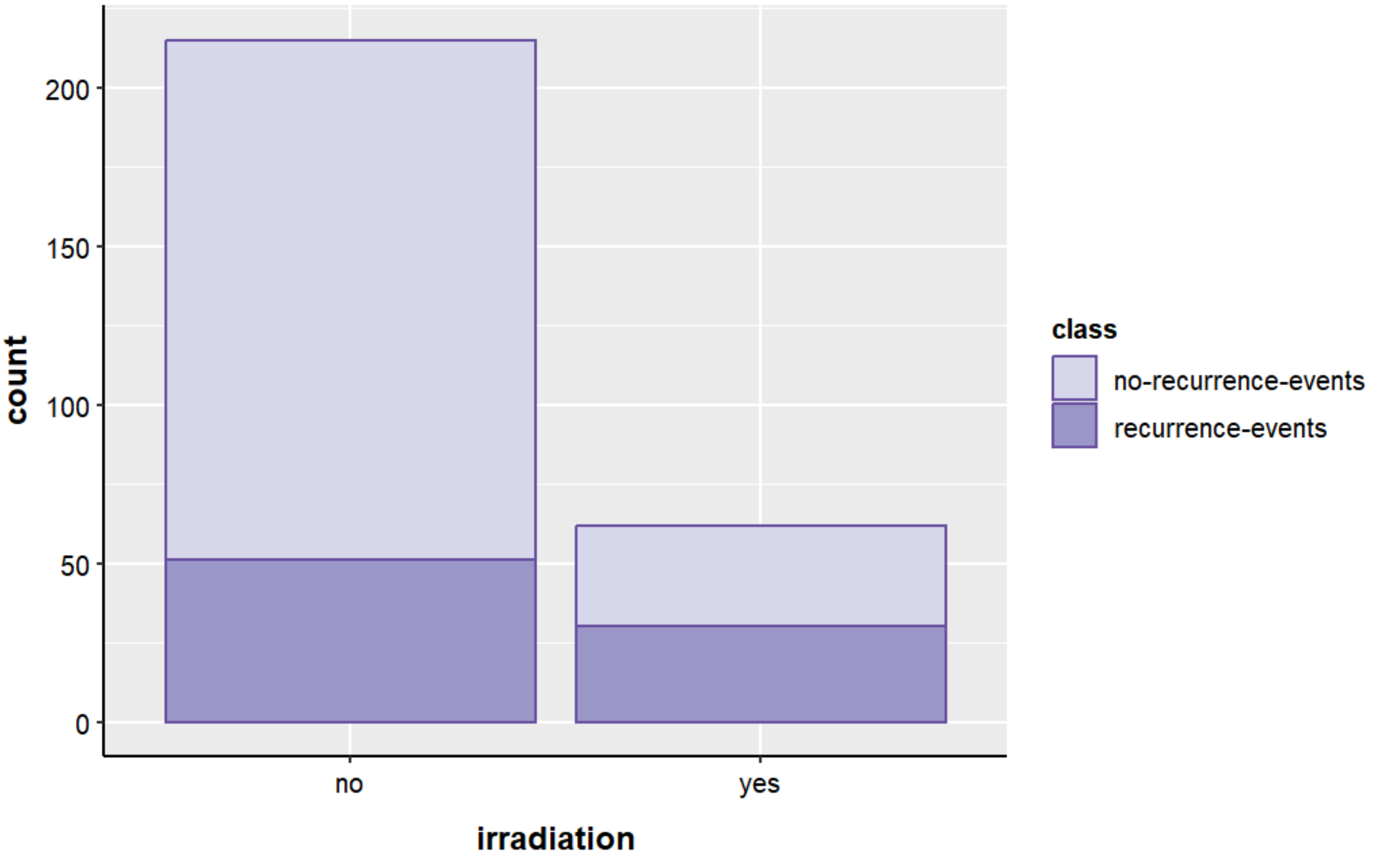

count
0
50
100
150
200
0-2
3-5
6-8
9-11
12-14
15-17
24-26
inv.nodes
class
no-recurrence-events
recurrence-events


count
0
50
100
150
200
no
yes
node.caps
class
no-recurrence-events
recurrence-events


count
0
50
100
1
2
3
deg.malignancy
class
no-recurrence-events
recurrence-events


150
100
50
0
count
left
right
breast
class
no-recurrence-events
recurrence-events


100
75
50
25
0
count
central
left_low
left_up
right_low
right_up
breast.quad
class
no-recurrence-events
recurrence-events


count
200
150
100
50
0
no
yes
irradiation
class
no-recurrence-events
recurrence-events