Math 308: Breast Cancer Bivariate Correspondence and Multiple Correspondence Analysis

Lily Samuel

2022-04-05

```
library(knitr)
library(systemfonts)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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
```

```
library(ggpubr )
library(FactoMineR)
library(reshape)
```

```
##
## Attaching package: 'reshape'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
##
## The following object is masked from 'package:dplyr':
##
##     rename
##
## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
```

```
library(readr)
library(kableExtra)
```

```
##
```

```
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
breast_cancer <- read_csv("/Users/lilysamuel/Desktop/breast_cancer_data.csv")
```

```
## New names:
## Rows: 285 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (9): no-recurrence-events, 30-39, premeno, 30-34, 0-2, no...6, left, lef... dbl
## (1): 3
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `no` -> `no...6`
## * `no` -> `no...10`
```

```
head(breast_cancer)
```

```
## # A tibble: 6 x 10
##   `no-recurrence-events` `30-39` premeno `30-34` `0-2` no...6   `3` left
##   <chr>                  <chr>   <chr>   <chr>   <chr> <chr>  <dbl> <chr>
## 1 no-recurrence-events   40-49   premeno 20-24   0-2   no         2 right
## 2 no-recurrence-events   40-49   premeno 20-24   0-2   no         2 left
## 3 no-recurrence-events   60-69   ge40    15-19   0-2   no         2 right
## 4 no-recurrence-events   40-49   premeno 0-4     0-2   no         2 right
## 5 no-recurrence-events   60-69   ge40    15-19   0-2   no         2 left
## 6 no-recurrence-events   50-59   premeno 25-29   0-2   no         2 left
## # i 2 more variables: left_low <chr>, no...10 <chr>
```

```
names(breast_cancer)<-c("RecEv","AgeGrp","Meno","Size","InvNodes",
                        "NodeCaps","DegMal","Side","Quad","Irrad")
breast_cancer<-breast_cancer %>%
  filter(Quad!="?",NodeCaps!="?")

breast_cancer<-breast_cancer %>%
  mutate(AgeGrp=paste("AG",AgeGrp,sep=""),
         InvNodes=paste("IN",InvNodes,sep=""))

head(breast_cancer) %>% kable(.) %>% kable_styling()
```

Task 1:

create a two-way contingency table of counts for this data.

Solution:

RecEv	AgeGrp	Meno	Size	InvNodes	NodeCaps	DegMal	Side	Quad	Irrad
no-recurrence-events	AG40-49	premeno	20-24	IN0-2	no	2	right	right_up	no
no-recurrence-events	AG40-49	premeno	20-24	IN0-2	no	2	left	left_low	no
no-recurrence-events	AG60-69	ge40	15-19	IN0-2	no	2	right	left_up	no
no-recurrence-events	AG40-49	premeno	0-4	IN0-2	no	2	right	right_low	no
no-recurrence-events	AG60-69	ge40	15-19	IN0-2	no	2	left	left_low	no
no-recurrence-events	AG50-59	premeno	25-29	IN0-2	no	2	left	left_low	no

```
part_one<-breast_cancer %>% select(Quad,DegMal)
head(part_one)
```

```
## # A tibble: 6 x 2
##   Quad      DegMal
##   <chr>      <dbl>
## 1 right_up       2
## 2 left_low       2
## 3 left_up        2
## 4 right_low      2
## 5 left_low       2
## 6 left_low       2
```

Task 2:

Create balloon plots for the observed cell proportions and the expected cell proportions.

Do you find visual evidence of dependence between the quadrant and the degree of malignancy?

Solution:

```
xtabs(~Quad+DegMal,data=part_one)
```

```
##            DegMal
## Quad         1  2  3
##   central    7 10  4
##   left_low  25 50 30
##   left_up   20 43 31
##   right_low  7 10  6
##   right_up   7 16 10
```

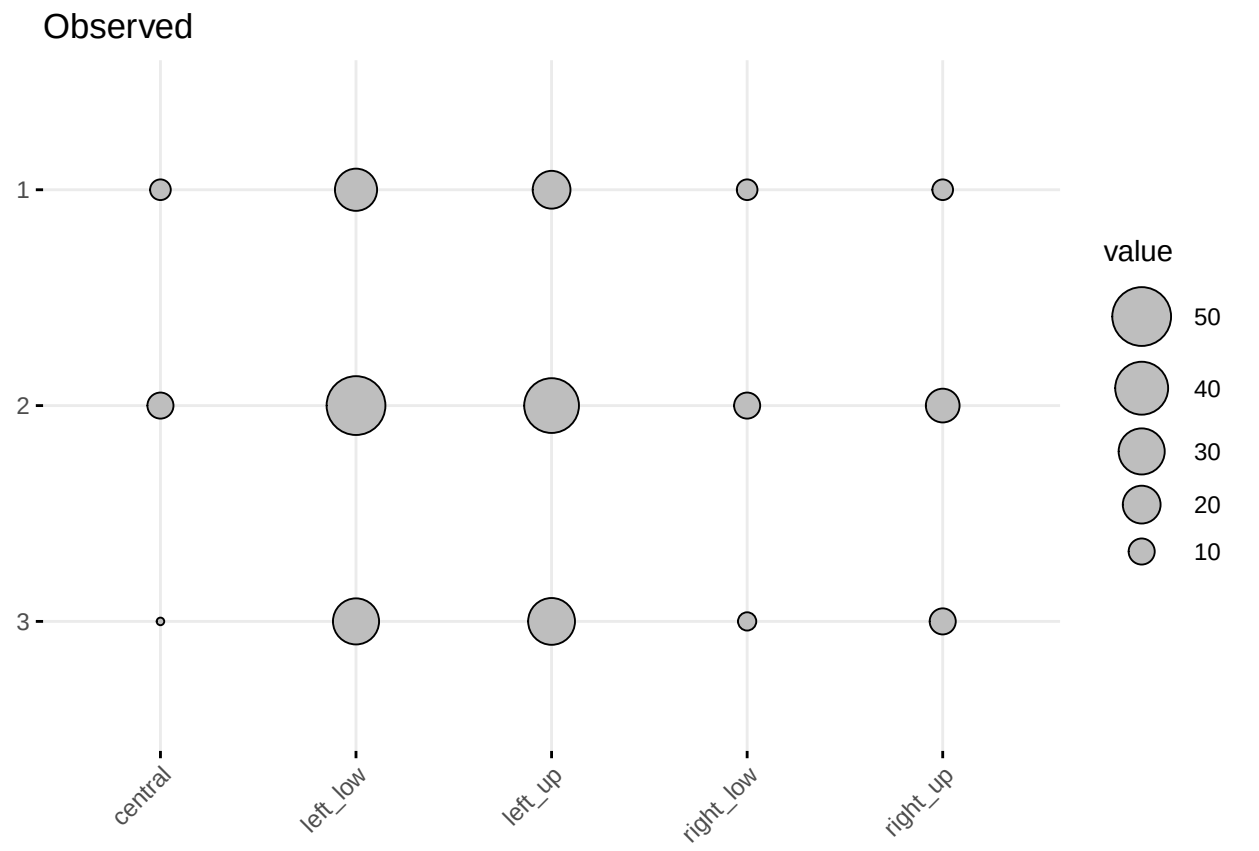
```
chisq_results<-chisq.test(xtabs(~Quad+DegMal,data=part_one))

obs<-melt(chisq_results$observed)
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by
## the caller; using TRUE

## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by
## the caller; using TRUE
```

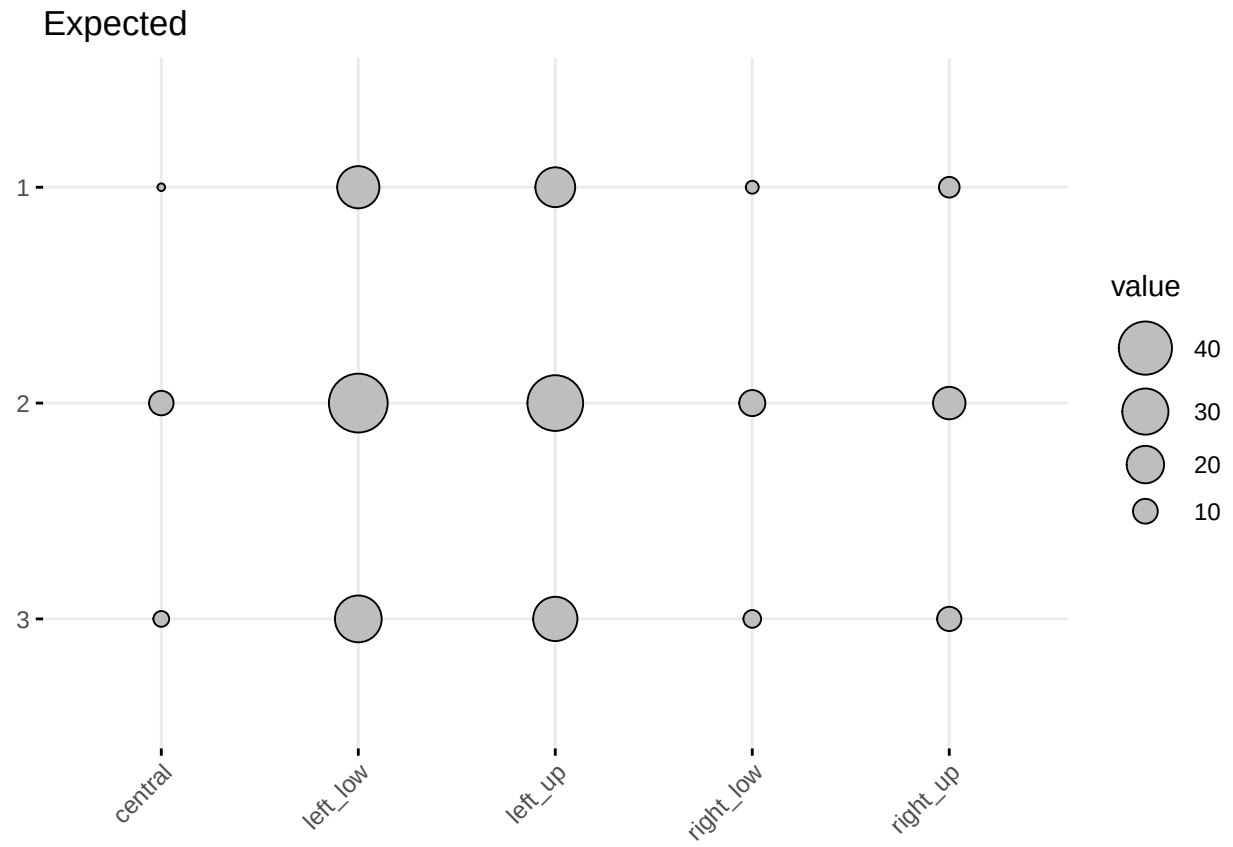
```
obs<-obs%>%mutate(DegMal=factor(DegMal))
exp<-melt(chisq_results$expected)
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by
## the caller; using TRUE

## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by
## the caller; using TRUE
```

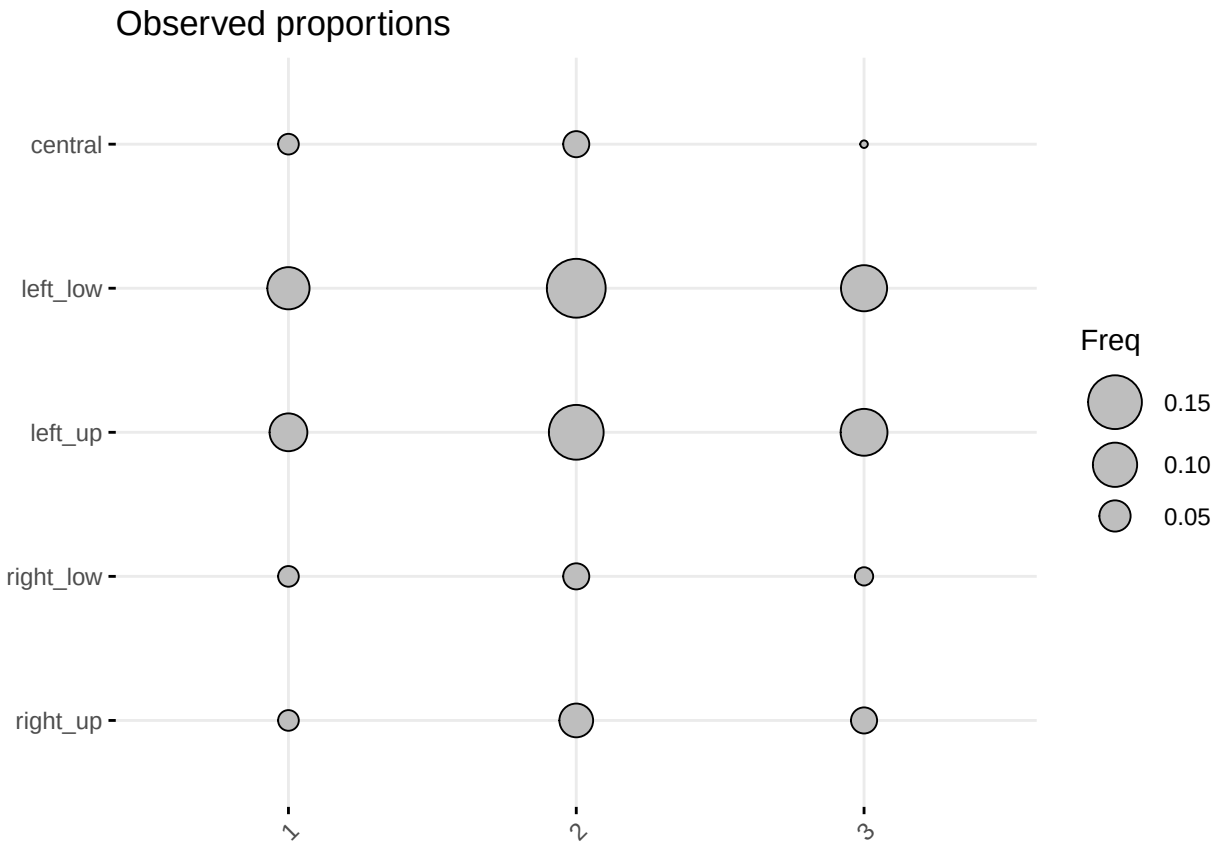
```
exp<-exp%>%mutate(DegMal=factor(DegMal))

ggballoonplot(obs, label=FALSE, show.margine=F, main="Observed")
```



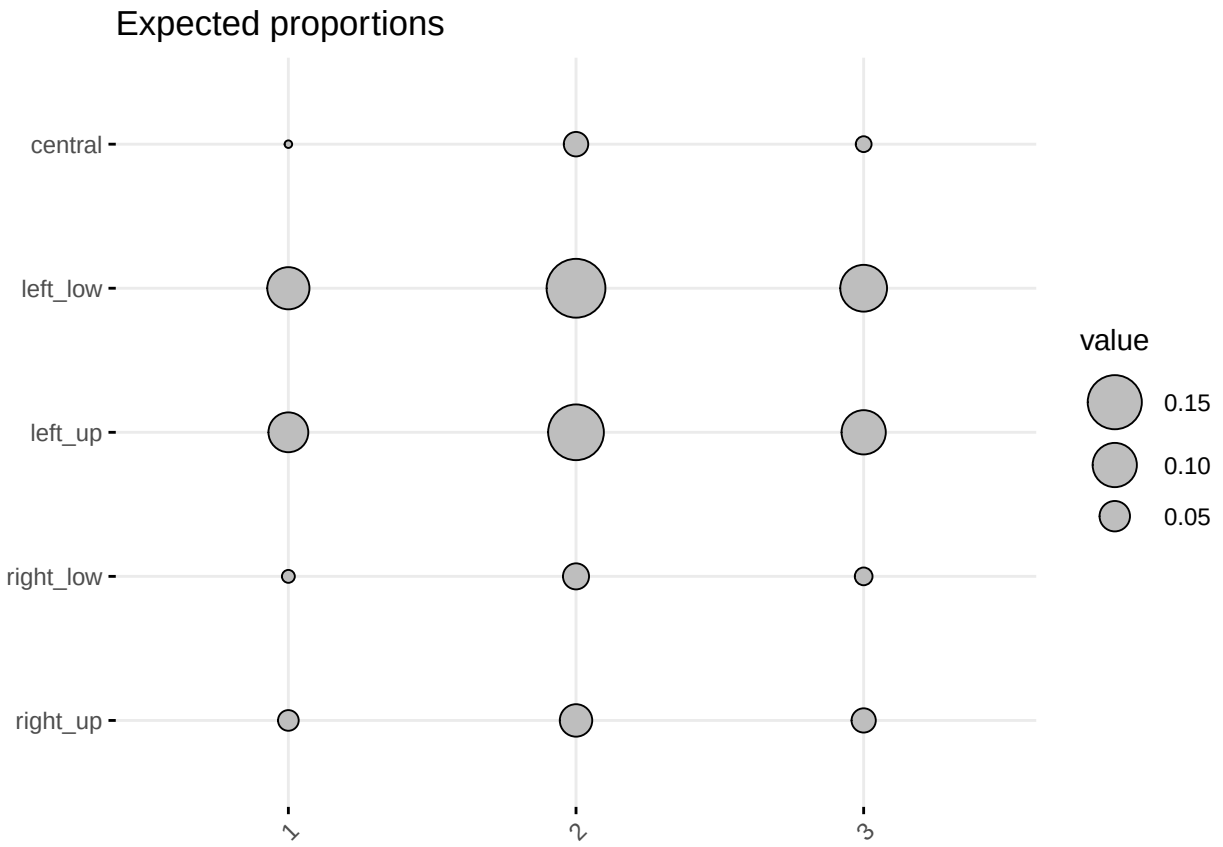

```
ggballoonplot(exp, label=FALSE, show.margine=F, main="Expected")
```



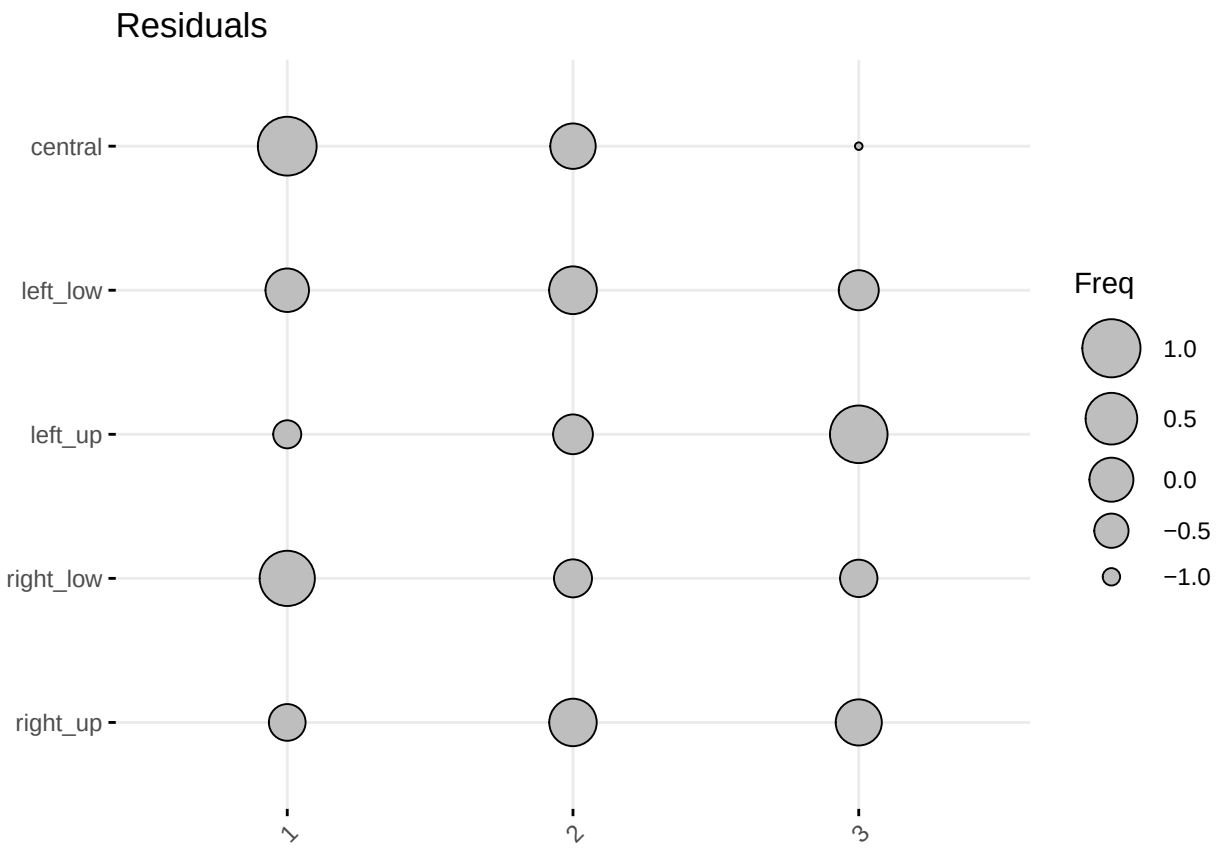

```
ggballoonplot(as.data.frame(t(chisq_results$observed/sum(chisq_results$observed)))) +
 ggtitle("Observed proportions")
```




```
ggballoonplot(as.data.frame(chisq_results$expected/sum(chisq_results$observed))) +
 ggtitle("Expected proportions")
```




```
ggballoonplot(as.data.frame(t(chisq_results$stdres))) +
 ggtitle("Residuals")
```



Based on what we see here, there is not a large amount of visual evidence of dependence. The residuals are also small, meaning only minor departures from dependence would be observed when doing the correspondence analysis.

Task 3:

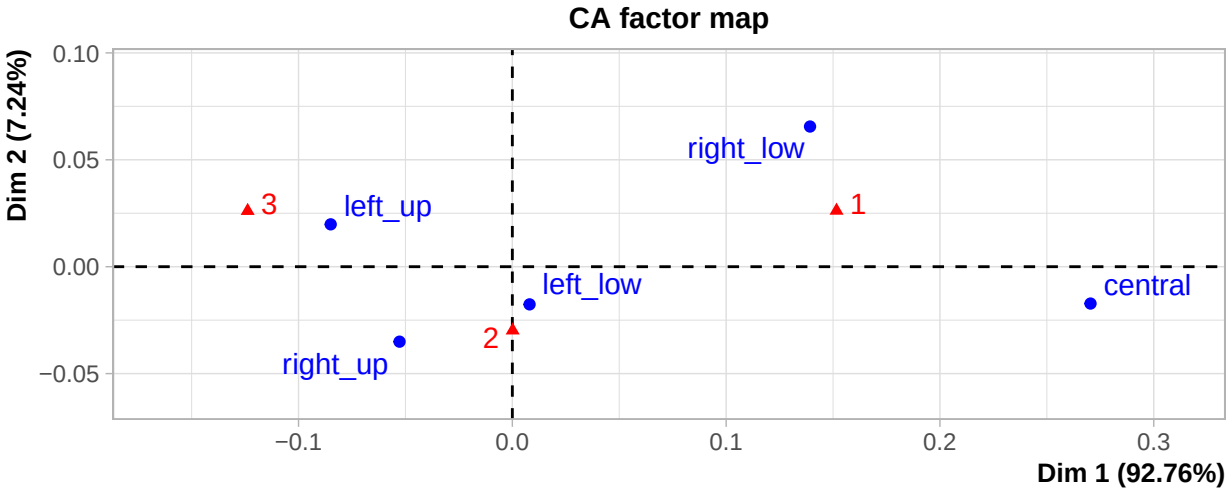
Conduct a correspondence analysis for this data, being sure to complete the following tasks:

- Report the table of eigenvalues for the 2 components and explain how many components you think are sufficient to analyze the data.
- Generate a row and column points factor map for the first two dimensions of the correspondence analysis (regardless of what your answer is to the first bullet point). Give separate interpretations of each dimension with respect to the associations between column and row points.

Solution:

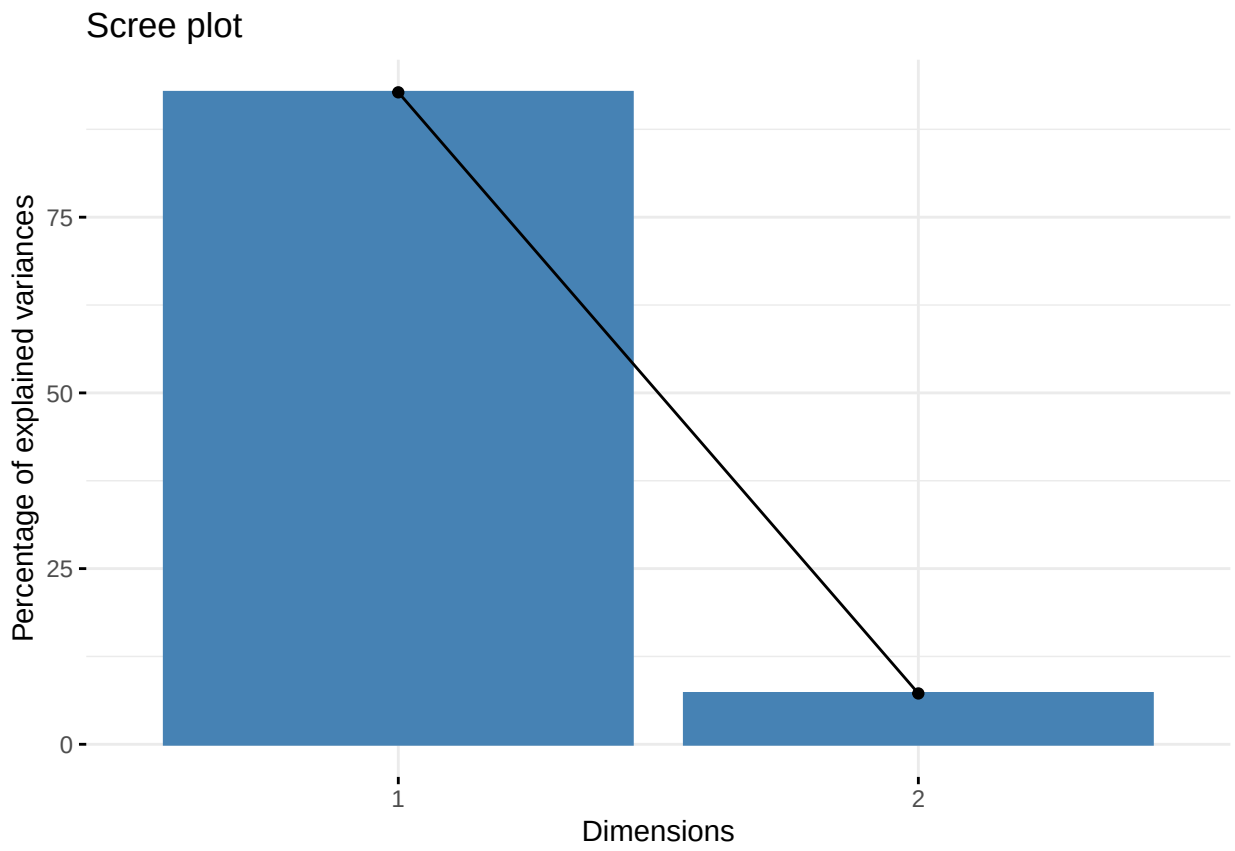
```
part_one_ac <- CA(xtabs(~Quad+DegMal,data=part_one))
```

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.0099956	92.764898	92.7649
dim 2	0.0007796	7.235102	100.0000

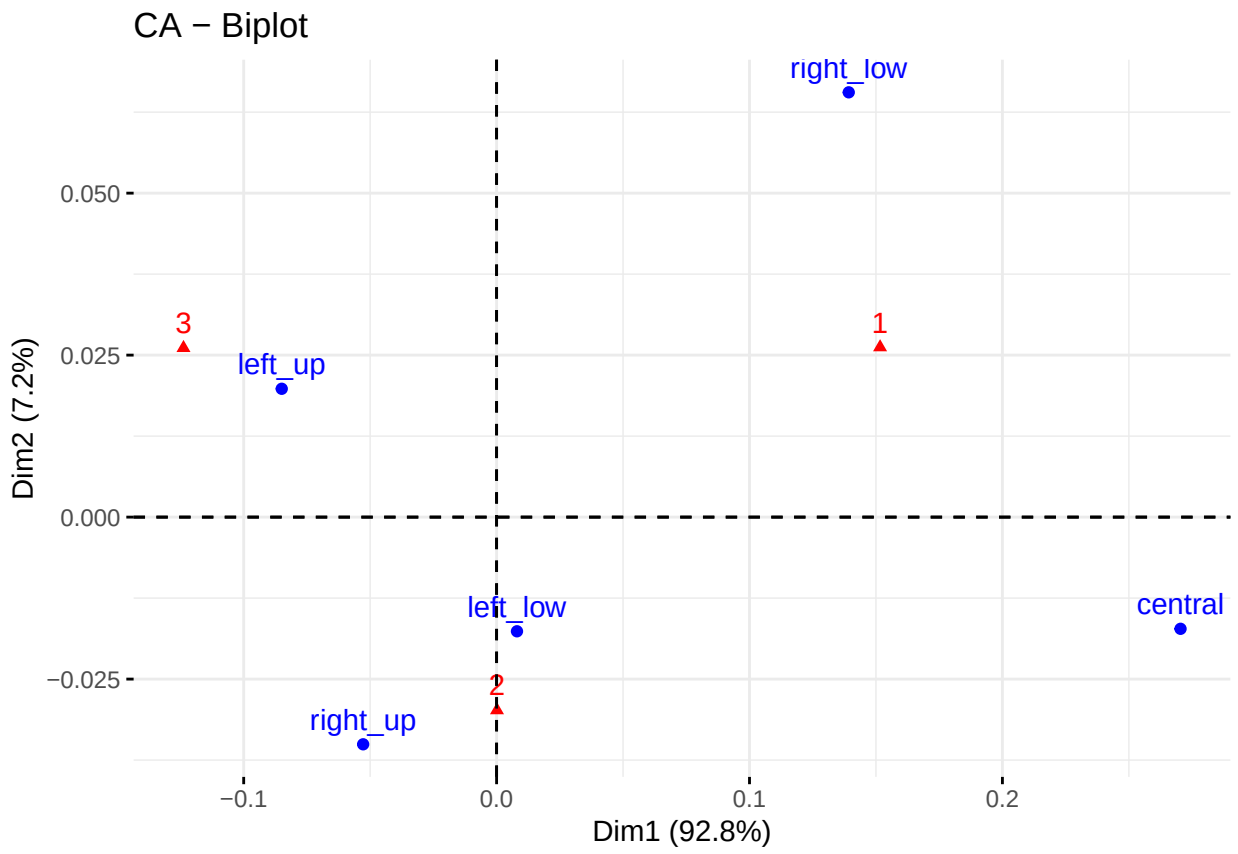



```
part_one_ac$eig %>% kable(.) %>% kable_styling()
```

```
fviz_eig(part_one_ac)
```




```
fviz_ca_biplot(part_one_ac)
```




```
part_one_ac$row
```

```
## $coord
##                  Dim 1       Dim 2
## central     0.270424239 -0.01724112
## left_low    0.008057385 -0.01762415
## left_up    -0.084959401  0.01980742
## right_low   0.139236054  0.06555987
## right_up   -0.052763028 -0.03506591
##
## $contrib
##                 Dim 1     Dim 2
## central    55.6661119  2.901147
## left_low    0.2470915 15.157402
## left_up    24.5941094 17.139685
## right_low 16.1626144 45.943419
## right_up   3.3300727 18.858347
##
## $cos2
##                Dim 1       Dim 2
## central    0.9959516 0.004048351
## left_low   0.1728786 0.827121373
## left_up    0.9484479 0.051552083
## right_low 0.8185292 0.181470845
## right_up   0.6936337 0.306366349
```

```
##
## $inertia
## [1] 0.0055868009 0.0001428654 0.0025919593 0.0019737312 0.0004798816
```

```
part_one_ac$col
```

```
## $coord
##           Dim 1       Dim 2
## 1  0.1516281345  0.02621789
## 2  0.0001558058 -0.02980569
## 3 -0.1237969856  0.02610559
##
## $contrib
##          Dim 1    Dim 2
## 1 5.500268e+01 21.08427
## 2 1.135109e-04 53.26076
## 3 4.499720e+01 25.65497
##
## $cos2
##          Dim 1      Dim 2
## 1 9.709704e-01 0.02902965
## 2 2.732482e-05 0.99997268
## 3 9.574253e-01 0.04257473
##
## $inertia
## [1] 0.0056622425 0.0004152319 0.0046977639
```

We see that we only need one component to analyze the data, as the cumulative percentage of variance in 93.6

The distance between any row or column points gives a measure of their similarity. Row points with similar profile are closed on the factor map. The same holds true for column points. Looking at the factor map, we can see that degree of malignancy is an equal distance in-between degree of malignancy 3 and degree of malignancy 2. It is evident that row category right_low contribute to the positive pole of the first dimension, while the category left up has a contribution to the negative pole of the first dimension. Left_low does not contribute to the first dimension, at is rests at O on the y axis. however it slightly contributes to the second dimension.

Task 4:

Conduct a correspondence analysis for this data, being sure to complete the following tasks: - Report the table of eigenvalues for the first 5 components and explain how many components you think are sufficient to analyze the data. - Generate a factor map for the first two dimensions of the correspondence analysis (regardless of what your answer is to the first bullet point). Give a summary of which levels of which variables are most strongly associated with each of the first two dimensions and how you made your decisions.

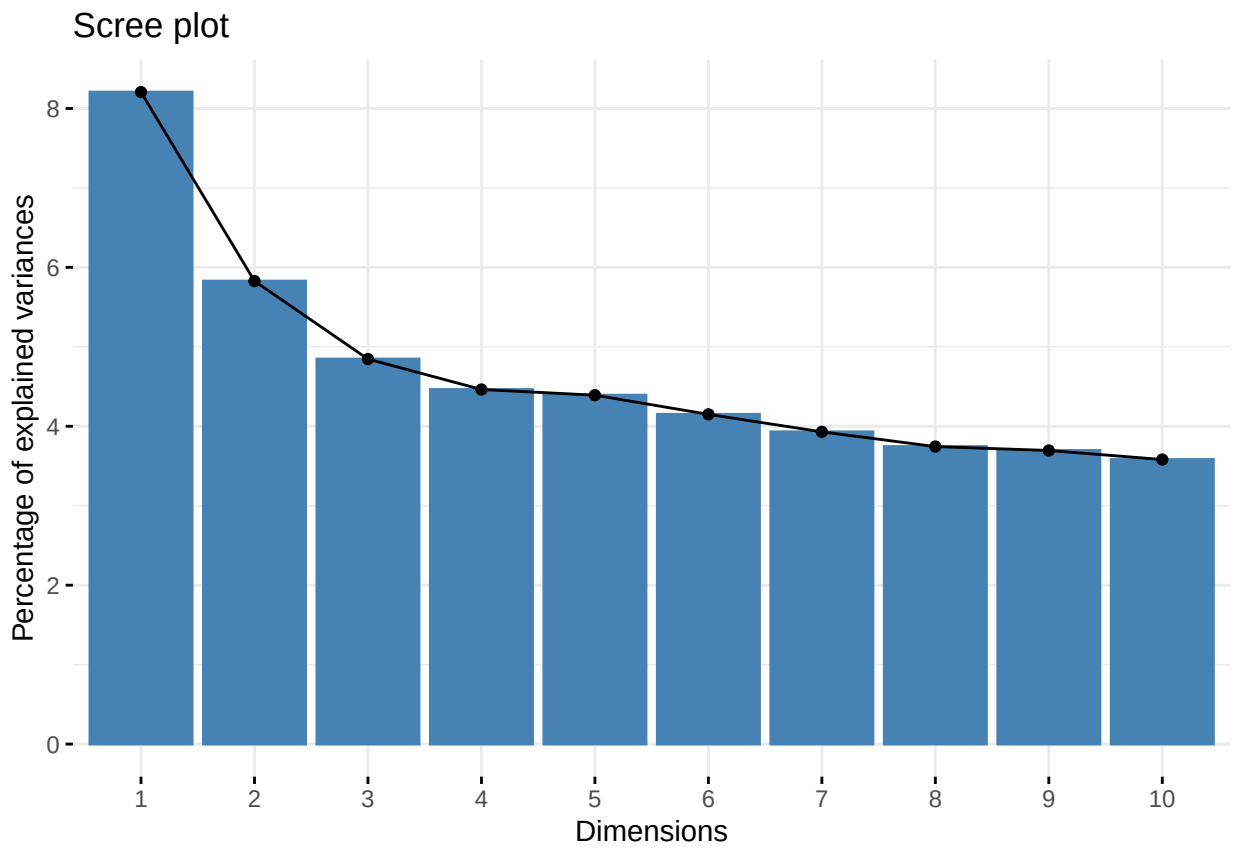
Solution:

```
breast_cancer<-breast_cancer %>% mutate_all(~factor(.))
```

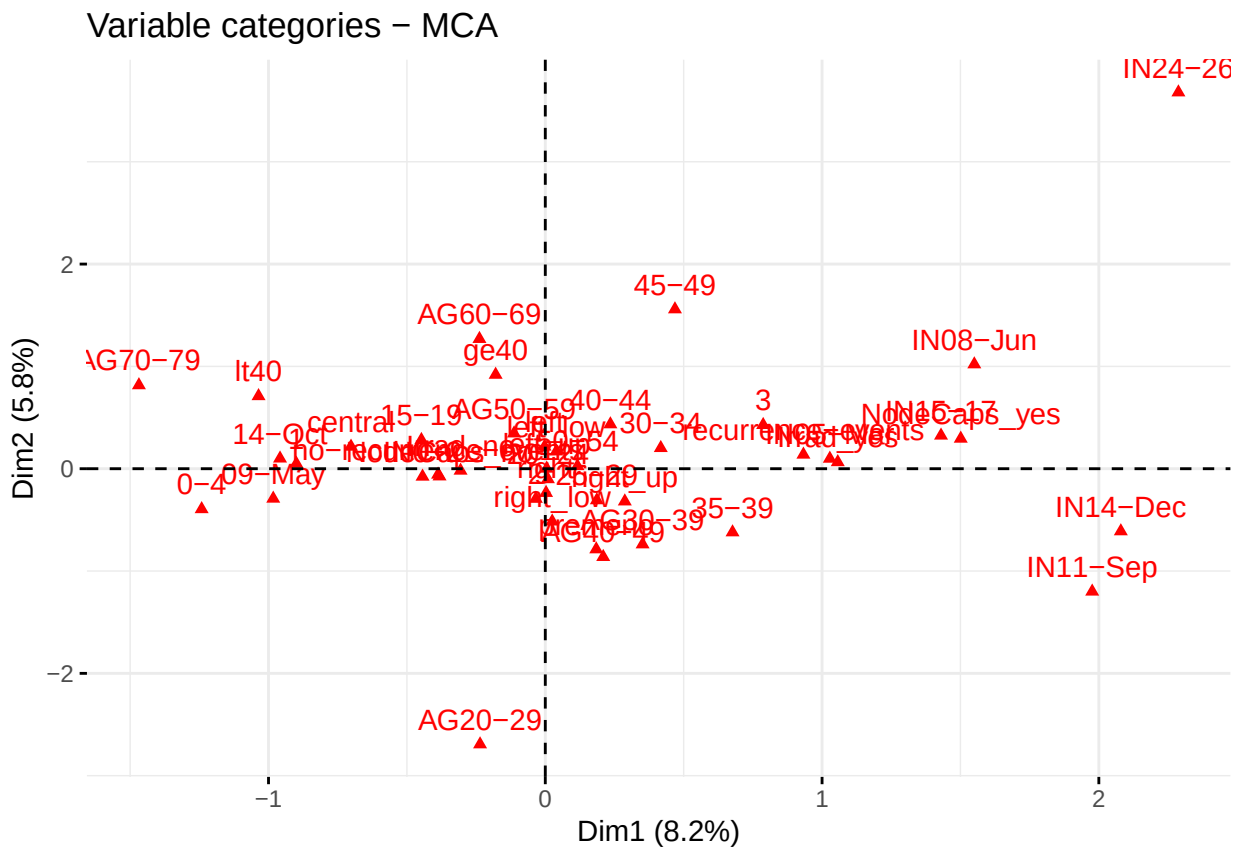
```
part_two_mca <- MCA(breast_cancer,graph=FALSE,ncp=5)
part_two_mca$eig %>% kable(.) %>% kable_styling()
```

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.2708011	8.2060954	8.206095
dim 2	0.1923003	5.8272811	14.033376
dim 3	0.1599240	4.8461826	18.879559
dim 4	0.1472714	4.4627712	23.342330
dim 5	0.1449115	4.3912568	27.733587
dim 6	0.1369558	4.1501748	31.883762
dim 7	0.1296869	3.9299047	35.813666
dim 8	0.1235632	3.7443386	39.558005
dim 9	0.1219386	3.6951091	43.253114
dim 10	0.1181864	3.5814049	46.834519
dim 11	0.1155654	3.5019829	50.336502
dim 12	0.1113492	3.3742182	53.710720
dim 13	0.1074218	3.2552053	56.965925
dim 14	0.1036165	3.1398933	60.105819
dim 15	0.1020368	3.0920239	63.197843
dim 16	0.0971969	2.9453619	66.143205
dim 17	0.0926116	2.8064118	68.949617
dim 18	0.0917445	2.7801358	71.729752
dim 19	0.0886795	2.6872582	74.417010
dim 20	0.0850935	2.5785903	76.995601
dim 21	0.0830696	2.5172611	79.512862
dim 22	0.0800748	2.4265086	81.939370
dim 23	0.0722253	2.1886440	84.128015
dim 24	0.0709812	2.1509442	86.278959
dim 25	0.0706192	2.1399758	88.418935
dim 26	0.0621217	1.8824755	90.301410
dim 27	0.0616526	1.8682604	92.169670
dim 28	0.0553691	1.6778508	93.847521
dim 29	0.0534561	1.6198828	95.467404
dim 30	0.0520023	1.5758265	97.043230
dim 31	0.0497211	1.5067014	98.549932
dim 32	0.0270073	0.8184019	99.368334
dim 33	0.0208450	0.6316664	100.000000

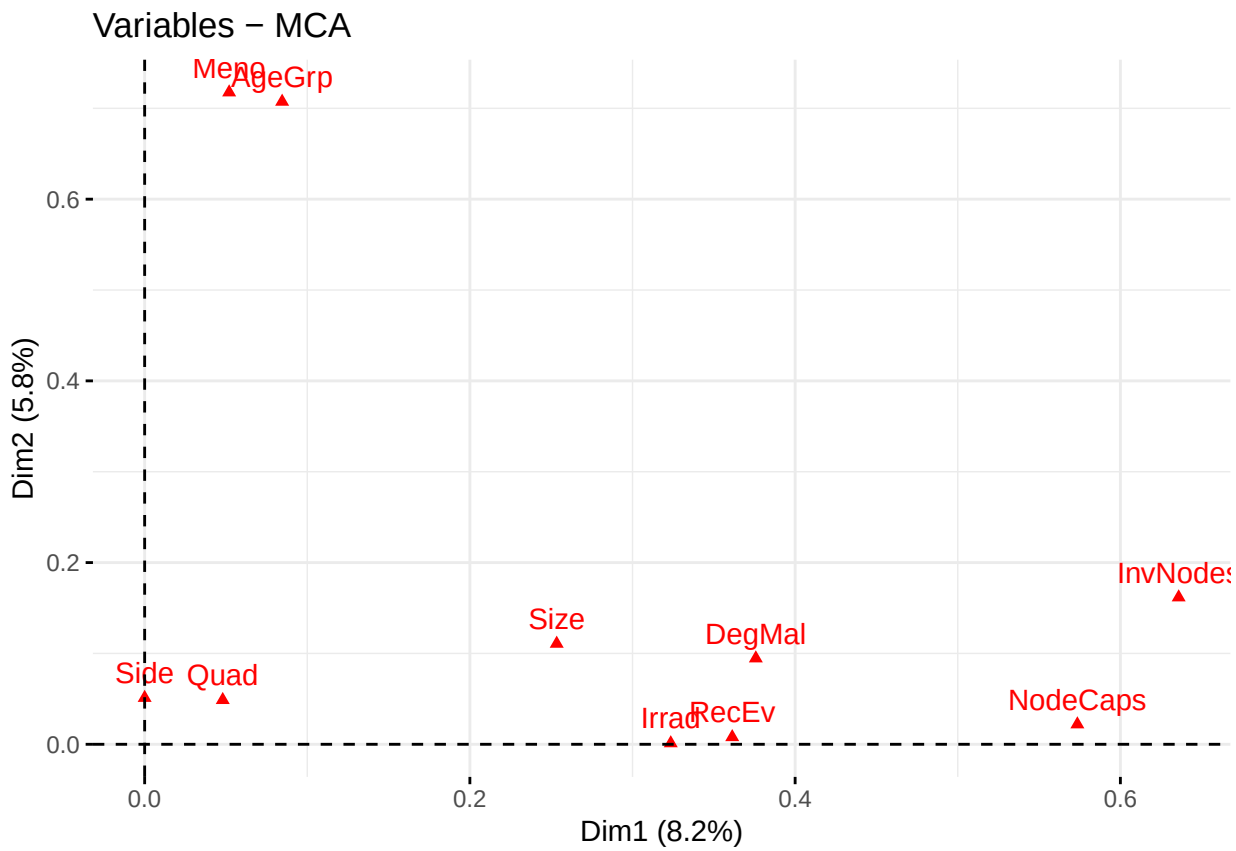
```
fviz_eig(part_two_mca)
```




```
fviz_mca_var(part_two_mca)
```




```
fviz_mca_var(part_two_mca,choice = "var")
```



We see that for the first dimension, InvNodes,NodeCaps,DegMal, Irrad, and Recurrent Events are the variables with levels that contribute the most.

With these variables, there visible contrasts between the opposite ends of the spectrum.

- InvNodes 0-2 is contrasted with all of the rest.

The second dimension is mostly driven by the Menopause and AgeGrp levels

- contrasting all groups 50 and over and gt40 menopausal with all groups 49 and younger and pre-menopausal.

Task 5:

```
part_two_mca_morecp <- MCA(breast_cancer,graph=FALSE,ncp=50)
sort(round(part_two_mca_morecp$var$eta2[1,],2),decreasing=TRUE)
```

```
##  Dim 1 Dim 27 Dim 14 Dim 31 Dim 10 Dim 12 Dim 15 Dim 22 Dim 26 Dim 29 Dim 20
##   0.36   0.12   0.10   0.10   0.04   0.04   0.03   0.03   0.03   0.03   0.02
## Dim 23  Dim 2  Dim 3 Dim 11 Dim 13 Dim 17 Dim 18 Dim 19  Dim 4  Dim 5  Dim 6
##   0.02   0.01   0.01   0.01   0.01   0.01   0.01   0.01   0.00   0.00   0.00
##  Dim 7  Dim 8  Dim 9 Dim 16 Dim 21 Dim 24 Dim 25 Dim 28 Dim 30 Dim 32 Dim 33
##   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
```

Recurrence events is only well represented by the first dimension.

The InvNodes,NodeCaps,DegMal and Irrad are the most strongly associated to recurrent events. We know this because the next largest eta2 value is 0.14 and it only appears on dimension 27.

We see that the more serious disease values are all positive along (RecurrentEvents=yes), and corresponding negative values for the rest (RecurrentEvents=no).