Chapter 5

Hanna Haltia

2023-12-05

Dimensionality reduction

Let's begin by loading in the data and exploring the dimensions.

```
# Load packages
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
```

```
# Load in data
human <- read.csv("data/human.csv")
str(human)
```

```
## 'data.frame':    155 obs. of  9 variables:
##  $ country  : chr  "Norway" "Australia" "Switzerland" "Denmark" ...
##  $ Edu2.FM  : num  1.007 0.997 0.983 0.989 0.969 ...
##  $ Labo.FM  : num  0.891 0.819 0.825 0.884 0.829 ...
##  $ Edu.Exp  : num  17.5 20.2 15.8 18.7 17.9 16.5 18.6 16.5 15.9 19.2 ...
##  $ Life.Exp : num  81.6 82.4 83 80.2 81.6 80.9 80.9 79.1 82 81.8 ...
##  $ GNI      : int  64992 42261 56431 44025 45435 43919 39568 52947 42155 32689 ...
##  $ Mat.Mor  : int  4 6 6 5 6 7 9 28 11 8 ...
##  $ Ado.Birth: num  7.8 12.1 1.9 5.1 6.2 3.8 8.2 31 14.5 25.3 ...
##  $ Parli.F  : num  39.6 30.5 28.5 38 36.9 36.9 19.9 19.4 28.2 31.4 ...
```

```
# Set "country" as row names
human_ <- column_to_rownames(human, "country")
```

Next let's explore the variables and their relations

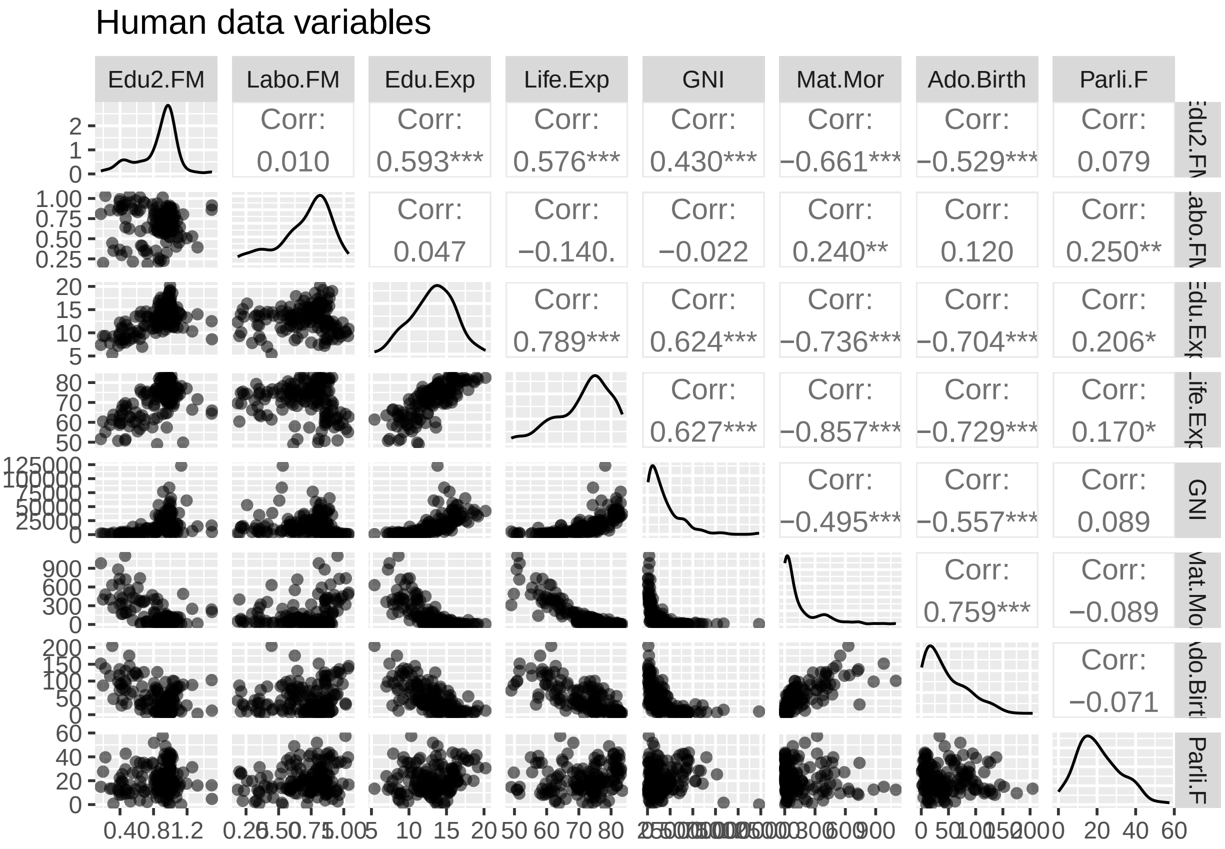
```
# Summaries of the variables
summary(human_)
```

```
##     Edu2.FM          Labo.FM          Edu.Exp         Life.Exp    
##  Min.   :0.1717   Min.   :0.1857   Min.   : 5.40   Min.   :49.00  
##  1st Qu.:0.7264   1st Qu.:0.5984   1st Qu.:11.25   1st Qu.:66.30  
##  Median :0.9375   Median :0.7535   Median :13.50   Median :74.20  
##  Mean   :0.8529   Mean   :0.7074   Mean   :13.18   Mean   :71.65  
##  3rd Qu.:0.9968   3rd Qu.:0.8535   3rd Qu.:15.20   3rd Qu.:77.25  
##  Max.   :1.4967   Max.   :1.0380   Max.   :20.20   Max.   :83.50  
##       GNI            Mat.Mor         Ado.Birth         Parli.F     
##  Min.   :   581   Min.   :   1.0   Min.   :  0.60   Min.   : 0.00  
##  1st Qu.:  4198   1st Qu.:  11.5   1st Qu.: 12.65   1st Qu.:12.40  
##  Median : 12040   Median :  49.0   Median : 33.60   Median :19.30  
##  Mean   : 17628   Mean   : 149.1   Mean   : 47.16   Mean   :20.91  
##  3rd Qu.: 24512   3rd Qu.: 190.0   3rd Qu.: 71.95   3rd Qu.:27.95  
##  Max.   :123124   Max.   :1100.0   Max.   :204.80   Max.   :57.50
```

```
# Plot matrix
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from   
##   +.gg   ggplot2
```

```
ggpairs(human_, mapping = aes(alpha=0.2),
              title="Human data variables",
              lower = list(combo = wrap("facethist", bins = 20)))
```

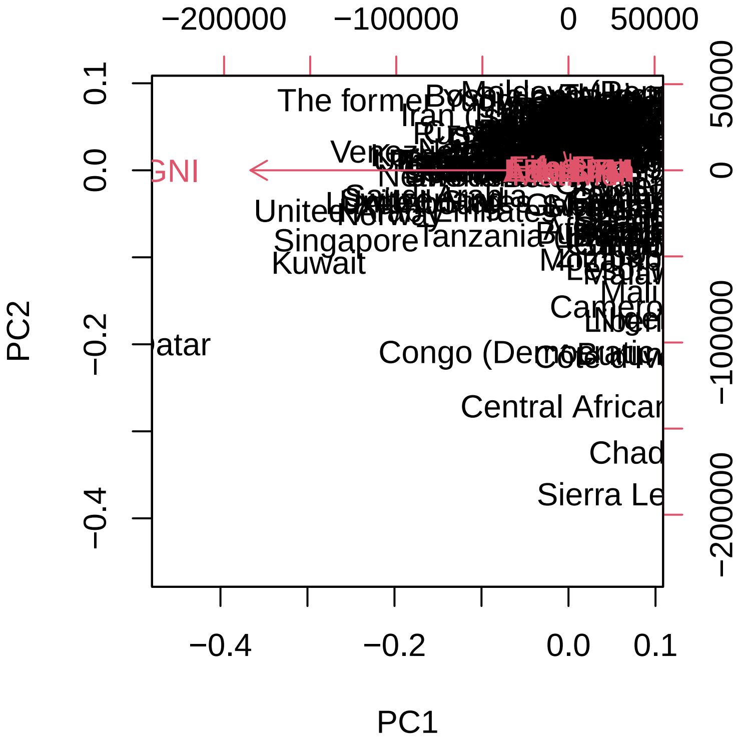


All variables in the dataset are now numerical as the country information has been moved to rownames. The distributions and ranges of the variables are not equal. Some pretty strong inverse correlations are observed between maternal mortality and life expectancy at birth, and adolescent birth rate and life expectancy at birth for example.

Next let's do PCA on the raw human data.

```r
# Perform principal component analysis (with the SVD method)
pca_human <- prcomp(human_)

# Draw a biplot of the principal component representation and the original variables
biplot(pca_human, choices = 1:2)
```

```
## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L], length =
## arrow.len): zero-length arrow is of indeterminate angle and so skipped
```

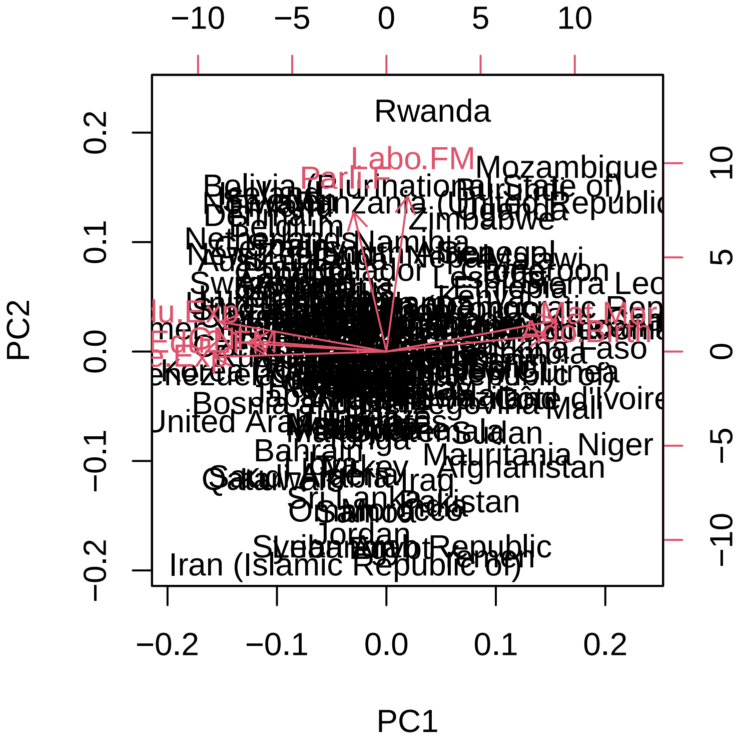


All data is on the upper right corner and only one PC vector is visible. The variation in the data is explained by GNI according to this plot.

Next let's standardize the data and perform the PCA again.

```r
# Standardize the variables
human_std <- scale(human_)

# Perform principal component analysis (with the SVD method)
pca_human <- prcomp(human_std)
```

```r
# Draw a biplot of the principal component representation and the original variables
biplot(pca_human, choices = 1:2)
```



```r
# Summary of the PCA
summary(pca_human)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.0708 1.1397 0.87505 0.77886 0.66196 0.53631 0.45900
## Proportion of Variance 0.5361 0.1624 0.09571 0.07583 0.05477 0.03595 0.02634
## Cumulative Proportion  0.5361 0.6984 0.79413 0.86996 0.92473 0.96069 0.98702
##                            PC8
## Standard deviation     0.32224
## Proportion of Variance 0.01298
## Cumulative Proportion  1.00000
```

Now the data is better spread and not all in the same corner. Also more vectors are visible dividing explaining the variance across several variables. Overall, it looks like the data needs scaling to be properly analyzed with PCA.

I think the results are different because when the data is not scaled, variables with bigger values get overestimated in their influence, since the variance in variables with smaller values is minor compared to the big values. GNI is a variable with huge numeric values compared to the others without scaling, so it becomes very influential in the first analysis.

Based on the table, PC1 explains about 53% of the variation and is affected by maternal mortality, adolescent birth rate, GNI, life expectancy at birth, and expected years of schooling at least. PC2 explains about 16% of the variation and is affected by percentage of females in parliament and ratio of females to males in the labour force.

Let’s move into the tea dataset.

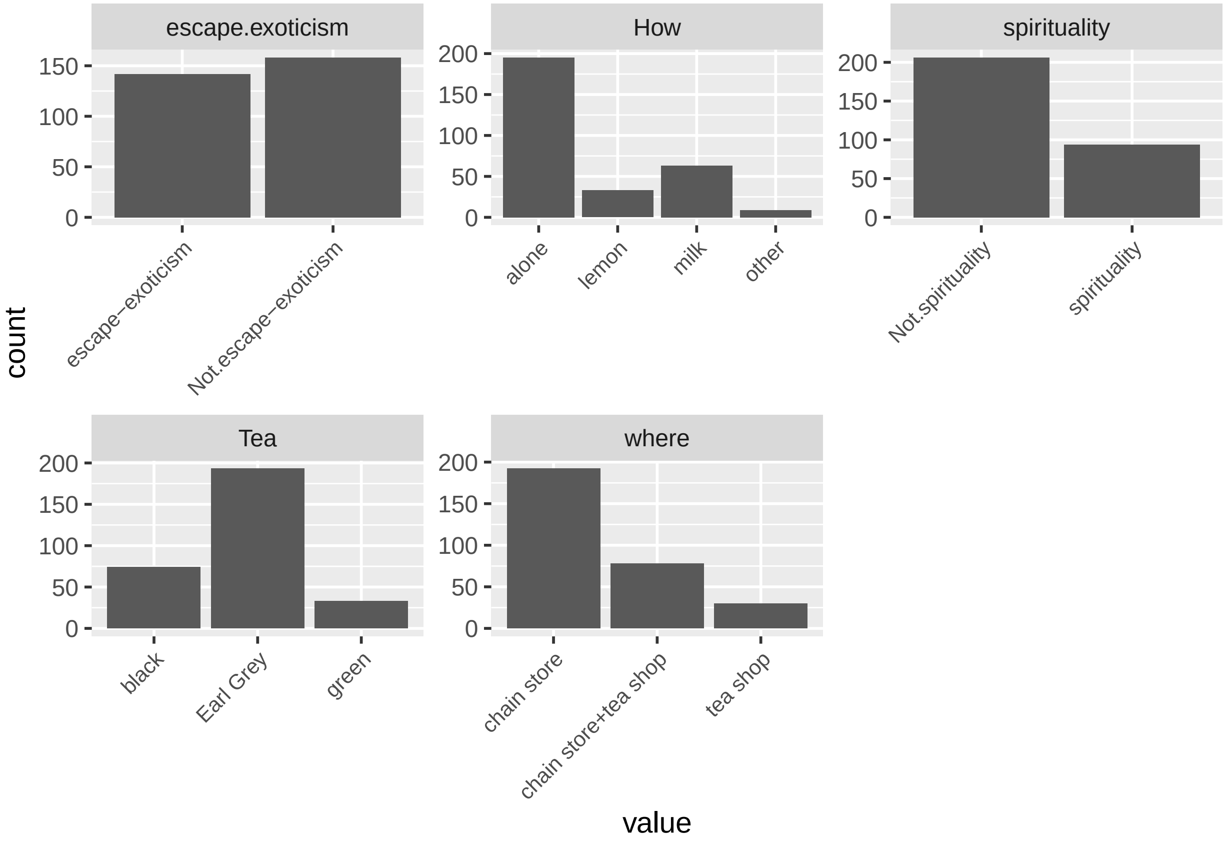
```r
# Load in the data
library(FactoMineR)
library(tidyverse)
tea <- read.csv("https://raw.githubusercontent.com/KimmoVehkalahti/Helsinki-Open-Data-Science/master/da

# Column names to keep in the dataset
keepers <- c("Tea", "How", "where", "spirituality", "escape.exoticism")

# Select the 'keep_columns' to create a new dataset
tea_time <- select(tea, all_of(keepers))

# look at the summaries and structure of the data
summary(tea_time)
```

```
##         Tea          How                        where               spirituality
##  black    : 74   alone:195   chain store         :192   Not.spirituality:206
##  Earl Grey:193   lemon: 33   chain store+tea shop: 78   spirituality    : 94
##  green    : 33   milk : 63   tea shop            : 30
##                  other:  9
##              escape.exoticism
##  escape-exoticism    :142
##  Not.escape-exoticism:158
##
##
```

```r
str(tea_time)
```

```
## 'data.frame':    300 obs. of  5 variables:
##  $ Tea             : Factor w/ 3 levels "black","Earl Grey",..: 1 1 2 2 2 2 2 1 2 1 ...
##  $ How             : Factor w/ 4 levels "alone","lemon",..: 1 3 1 1 1 1 1 3 3 1 ...
##  $ where           : Factor w/ 3 levels "chain store",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ spirituality    : Factor w/ 2 levels "Not.spirituality",..: 1 1 1 2 2 1 1 1 1 1 ...
##  $ escape.exoticism: Factor w/ 2 levels "escape-exoticism",..: 2 1 2 1 1 2 2 2 2 2 ...
```

```r
# visualize the dataset
library(ggplot2)
pivot_longer(tea_time, cols = everything()) %>%
  ggplot(aes(value)) + facet_wrap("name", scales = "free") +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))
```



From the chosen variables can be seen that most people have no escape exoticism related to tea, drink without adding anything with no spiritual meaning. The tea is most commonly specifically Earl Grey bought from a chain store.

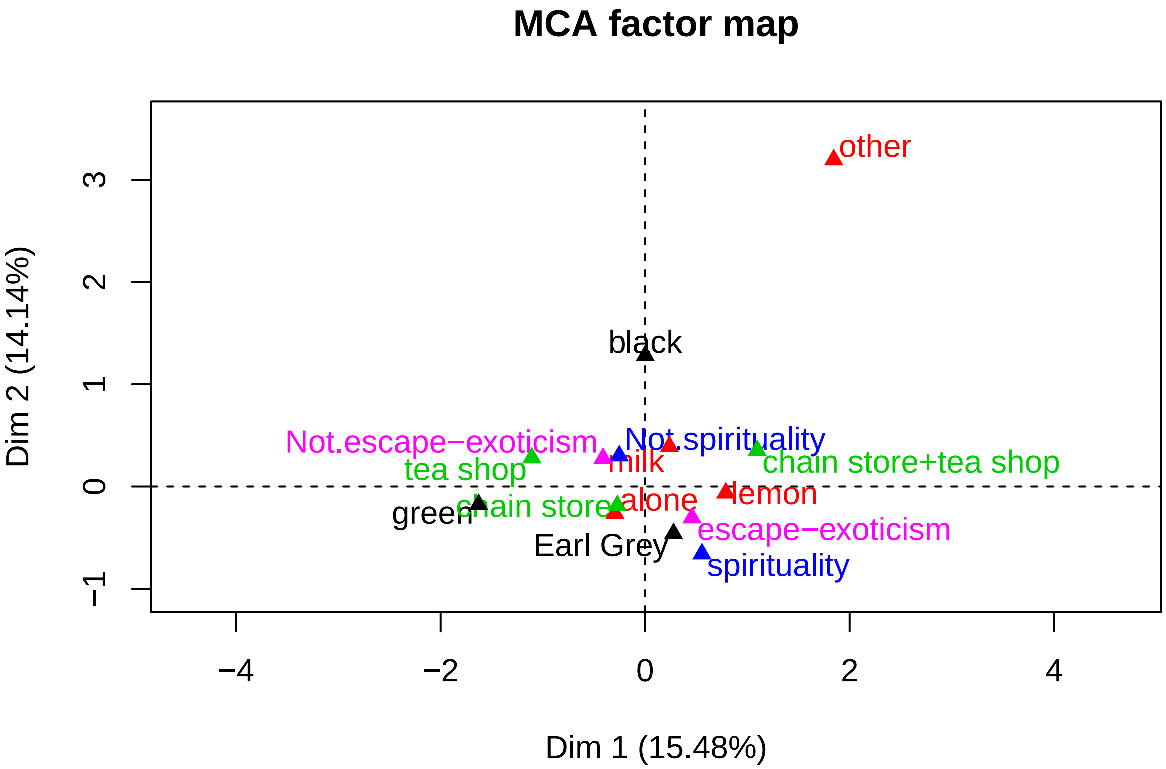
Next we will do MCA on the chosen variables of the tea dataset.

```
# Multiple correspondence analysis
library(FactoMineR)
mca <- MCA(tea_time, graph = FALSE)

# Summary of the model
summary(mca)
```

```
##
## Call:
## MCA(X = tea_time, graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               0.279   0.254   0.244   0.200   0.193   0.180   0.159
## % of var.             15.476  14.136  13.549  11.118  10.717  10.021   8.838
## Cumulative % of var.  15.476  29.612  43.161  54.279  64.995  75.016  83.854
##                        Dim.8   Dim.9
## Variance               0.149   0.142
## % of var.              8.260   7.887
## Cumulative % of var.  92.113 100.000
##
```

```
## Individuals (the 10 first)
##                         Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## 1                    | -0.467  0.261  0.198 |  0.557  0.407  0.282 | -0.354
## 2                    |  0.065  0.005  0.002 |  0.586  0.450  0.192 | -0.602
## 3                    | -0.362  0.157  0.218 | -0.133  0.023  0.029 | -0.260
## 4                    |  0.274  0.090  0.076 | -0.743  0.724  0.557 | -0.250
## 5                    |  0.274  0.090  0.076 | -0.743  0.724  0.557 | -0.250
## 6                    | -0.362  0.157  0.218 | -0.133  0.023  0.029 | -0.260
## 7                    | -0.362  0.157  0.218 | -0.133  0.023  0.029 | -0.260
## 8                    | -0.265  0.084  0.040 |  0.816  0.873  0.382 | -0.430
## 9                    |  0.359  0.154  0.075 |  0.340  0.152  0.068 |  0.026
## 10                   |  0.052  0.003  0.002 |  0.771  0.779  0.381 |  0.008
##                         ctr   cos2
## 1                      0.171  0.114 |
## 2                      0.495  0.202 |
## 3                      0.093  0.113 |
## 4                      0.085  0.063 |
## 5                      0.085  0.063 |
## 6                      0.093  0.113 |
## 7                      0.093  0.113 |
## 8                      0.252  0.106 |
## 9                      0.001  0.000 |
## 10                     0.000  0.000 |
##
## Categories (the 10 first)
##                          Dim.1     ctr    cos2  v.test     Dim.2     ctr
## black                |   0.001   0.000   0.000   0.014 |   1.280  31.755
## Earl Grey            |   0.278   3.568   0.139   6.454 |  -0.461  10.728
## green                |  -1.629  20.948   0.328  -9.901 |  -0.176   0.268
## alone                |  -0.295   4.073   0.162  -6.962 |  -0.263   3.535
## lemon                |   0.789   4.917   0.077   4.797 |  -0.063   0.034
## milk                 |   0.238   0.851   0.015   2.119 |   0.391   2.519
## other                |   1.845   7.328   0.105   5.609 |   3.195  24.077
## chain store          |  -0.272   3.404   0.132  -6.275 |  -0.187   1.760
## chain store+tea shop |   1.096  22.426   0.422  11.234 |   0.353   2.541
## tea shop             |  -1.108   8.811   0.136  -6.385 |   0.280   0.618
##                         cos2  v.test     Dim.3     ctr    cos2  v.test
## black                  0.536  12.663 |  -0.261   1.381   0.022  -2.585 |
## Earl Grey              0.383 -10.697 |  -0.031   0.049   0.002  -0.711 |
## green                  0.004  -1.069 |   0.765   5.276   0.072   4.649 |
## alone                  0.129  -6.199 |  -0.194   1.998   0.070  -4.562 |
## lemon                  0.000  -0.382 |   1.927  33.505   0.459  11.716 |
## milk                   0.041   3.483 |  -0.381   2.496   0.039  -3.394 |
## other                  0.316   9.717 |  -0.207   0.105   0.001  -0.629 |
## chain store            0.062  -4.312 |  -0.479  12.019   0.407 -11.033 |
## chain store+tea shop   0.044   3.614 |   0.416   3.686   0.061   4.262 |
## tea shop               0.009   1.616 |   1.982  32.200   0.436  11.421 |
##
## Categorical variables (eta2)
##                        Dim.1 Dim.2 Dim.3
## Tea                  | 0.341 0.544 0.082 |
## How                  | 0.239 0.384 0.465 |
## where                | 0.482 0.063 0.584 |
## spirituality         | 0.141 0.198 0.044 |
## escape.exoticism     | 0.189 0.084 0.045 |
```

```
# Visualize MCA
plot(mca, invisible=c("ind"), graph.type = "classic", habillage = "quali")
```



Based on the MCA factor map plot, spirituality, Earl Grey, escape-exoticism, and lemon group together. Tea shop, no escape-exoticism, and green tea also somewhat group together. Chain store tea buyers seem to drink tea without adding anything. Drinking tea "other" (not alone but also no lemon or milk) way is not associated to any of the categories studied.