CHAPTER 5 Dimensionality reduction

#*TEA TIME* #Load the tea dataset from the package Factominer. Explore the data briefly: look at the structure and the dimensions of the data and visualize it. Then do Multiple Correspondence Analysis on the tea data (or to a certain columns of the data, it’s up to you). Interpret the results of the MCA and draw at least the variable biplot of the analysis. You can also explore other plotting options for MCA. Comment on the output of the plots. (0-4 points)

```
library(FactoMineR)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
data("tea")
str(tea)
```

```
## 'data.frame':    300 obs. of  36 variables:
##  $ breakfast       : Factor w/ 2 levels "breakfast","Not.breakfast": 1 1 2 2 1 2 1 2 1 1
...
##  $ tea.time        : Factor w/ 2 levels "Not.tea time",..: 1 1 2 1 1 1 2 2 2 1 ...
##  $ evening         : Factor w/ 2 levels "evening","Not.evening": 2 2 1 2 1 2 2 1 2 1 ...
##  $ lunch           : Factor w/ 2 levels "lunch","Not.lunch": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dinner          : Factor w/ 2 levels "dinner","Not.dinner": 2 2 1 1 2 1 2 2 2 2 ...
##  $ always          : Factor w/ 2 levels "always","Not.always": 2 2 2 2 1 2 2 2 2 2 ...
##  $ home            : Factor w/ 2 levels "home","Not.home": 1 1 1 1 1 1 1 1 1 1 ...
##  $ work            : Factor w/ 2 levels "Not.work","work": 1 1 2 1 1 1 1 1 1 1 ...
##  $ tearoom         : Factor w/ 2 levels "Not.tearoom",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ friends         : Factor w/ 2 levels "friends","Not.friends": 2 2 1 2 2 2 1 2 2 2 ...
##  $ resto           : Factor w/ 2 levels "Not.resto","resto": 1 1 2 1 1 1 1 1 1 1 ...
##  $ pub             : Factor w/ 2 levels "Not.pub","pub": 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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 ...
##  $ sugar           : Factor w/ 2 levels "No.sugar","sugar": 2 1 1 2 1 1 1 1 1 1 ...
##  $ how             : Factor w/ 3 levels "tea bag","tea bag+unpackaged",..: 1 1 1 1 1 1 1 1
2 2 ...
##  $ where           : Factor w/ 3 levels "chain store",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ price           : Factor w/ 6 levels "p_branded","p_cheap",..: 4 6 6 6 6 3 6 6 5 5 ...
##  $ age             : int  39 45 47 23 48 21 37 36 40 37 ...
##  $ sex             : Factor w/ 2 levels "F","M": 2 1 1 2 2 2 2 1 2 2 ...
##  $ SPC             : Factor w/ 7 levels "employee","middle",..: 2 2 4 6 1 6 5 2 5 5 ...
##  $ Sport           : Factor w/ 2 levels "Not.sportsman",..: 2 2 2 1 2 2 2 2 2 1 ...
##  $ age_Q           : Factor w/ 5 levels "15-24","25-34",..: 3 4 4 1 4 1 3 3 3 3 ...
##  $ frequency       : Factor w/ 4 levels "1/day","1 to 2/week",..: 1 1 3 1 3 1 4 2 3 3 ...
##  $ escape.exoticism: Factor w/ 2 levels "escape-exoticism",..: 2 1 2 1 1 2 2 2 2 2 ...
##  $ spirituality    : Factor w/ 2 levels "Not.spirituality",..: 1 1 1 2 2 1 1 1 1 1 ...
##  $ healthy         : Factor w/ 2 levels "healthy","Not.healthy": 1 1 1 1 2 1 1 1 2 1 ...
##  $ diuretic        : Factor w/ 2 levels "diuretic","Not.diuretic": 2 1 1 2 1 2 2 2 2 1 ...
##  $ friendliness    : Factor w/ 2 levels "friendliness",..: 2 2 1 2 1 2 2 1 2 1 ...
##  $ iron.absorption : Factor w/ 2 levels "iron absorption",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ feminine        : Factor w/ 2 levels "feminine","Not.feminine": 2 2 2 2 2 2 2 1 2 2 ...
##  $ sophisticated   : Factor w/ 2 levels "Not.sophisticated",..: 1 1 1 2 1 1 1 2 2 1 ...
##  $ slimming        : Factor w/ 2 levels "No.slimming",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ exciting        : Factor w/ 2 levels "exciting","No.exciting": 2 1 2 2 2 2 2 2 2 2 ...
##  $ relaxing        : Factor w/ 2 levels "No.relaxing",..: 1 1 2 2 2 2 2 2 2 2 ...
##  $ effect.on.health: Factor w/ 2 levels "effect on health",..: 2 2 2 2 2 2 2 2 2 2 ...
```

```
dim(tea)
```

```
## [1] 300  36
```

```
keep_columns <- c("Tea", "How", "how", "sugar", "where", "lunch")
tea_time <- select(tea, one_of(keep_columns))
summary(tea_time)
```

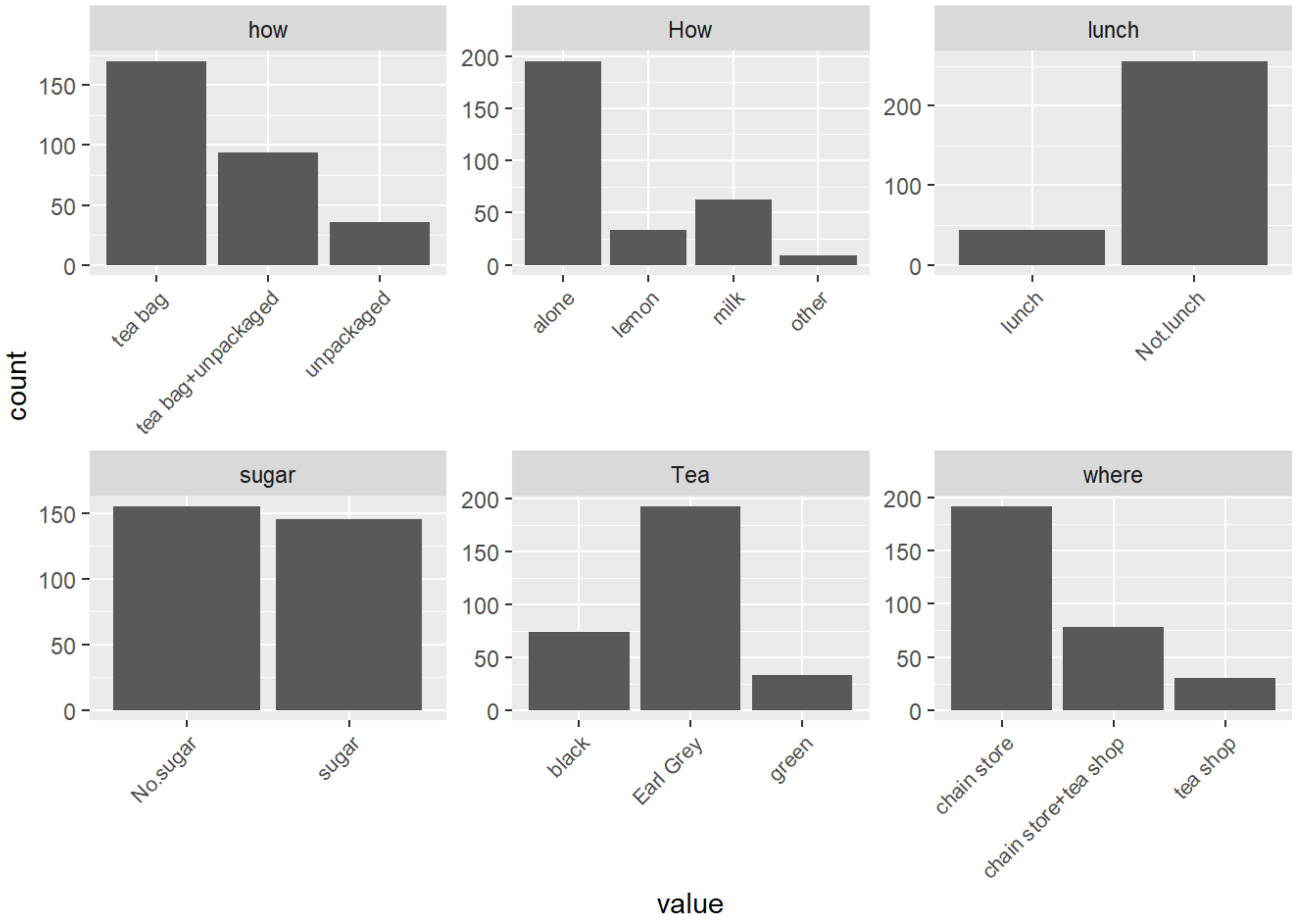
```
##         Tea         How                        how           sugar    
##  black    : 74   alone:195   tea bag           :170   No.sugar:155  
##  Earl Grey:193   lemon: 33   tea bag+unpackaged: 94   sugar   :145  
##  green    : 33   milk : 63   unpackaged        : 36                 
##                  other:  9                                          
##                   where          lunch    
##  chain store         :192   lunch    : 44  
##  chain store+tea shop: 78   Not.lunch:256  
##  tea shop            : 30                  
## 
```

```
str(tea_time)
```

```
## 'data.frame':    300 obs. of  6 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 ...
##  $ how  : Factor w/ 3 levels "tea bag","tea bag+unpackaged",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ sugar: Factor w/ 2 levels "No.sugar","sugar": 2 1 1 2 1 1 1 1 1 1 ...
##  $ where: Factor w/ 3 levels "chain store",..: 1 1 1 1 1 1 1 1 2 2 ...
##  $ lunch: Factor w/ 2 levels "lunch","Not.lunch": 2 2 2 2 2 2 2 2 2 2 ...
```

```
gather(tea_time) %>% ggplot(aes(value)) + facet_wrap("key", scales = "free") + geom_bar() + t
heme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))
```

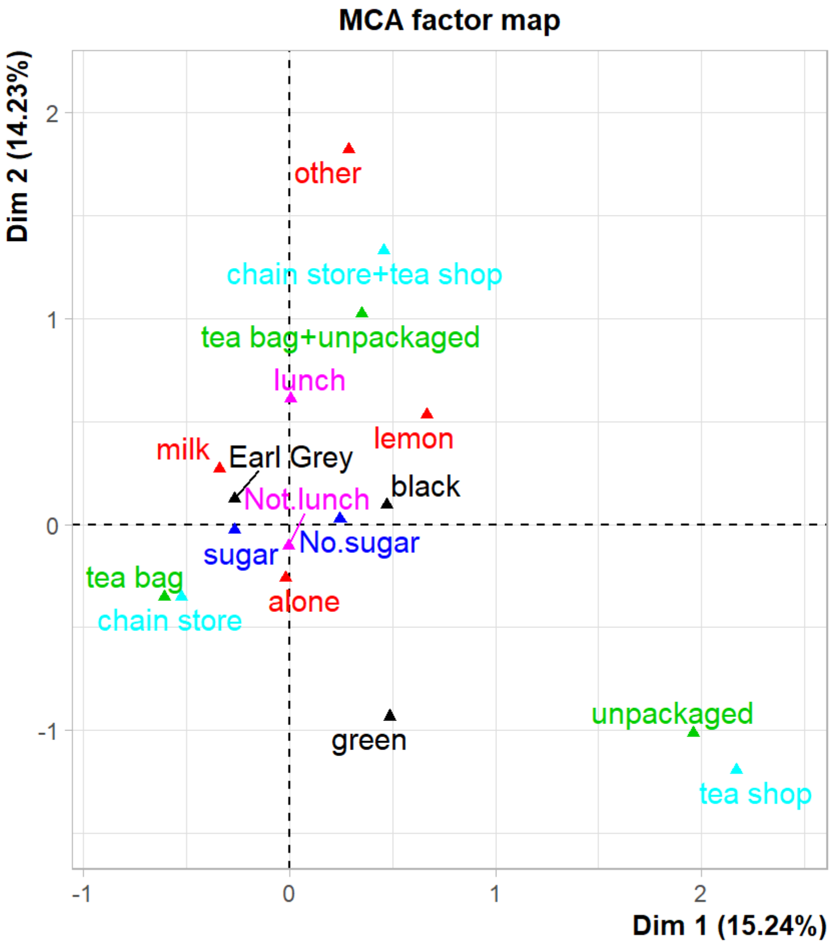
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```



```
mca <- MCA(tea_time, graph = FALSE)
summary(mca)
```

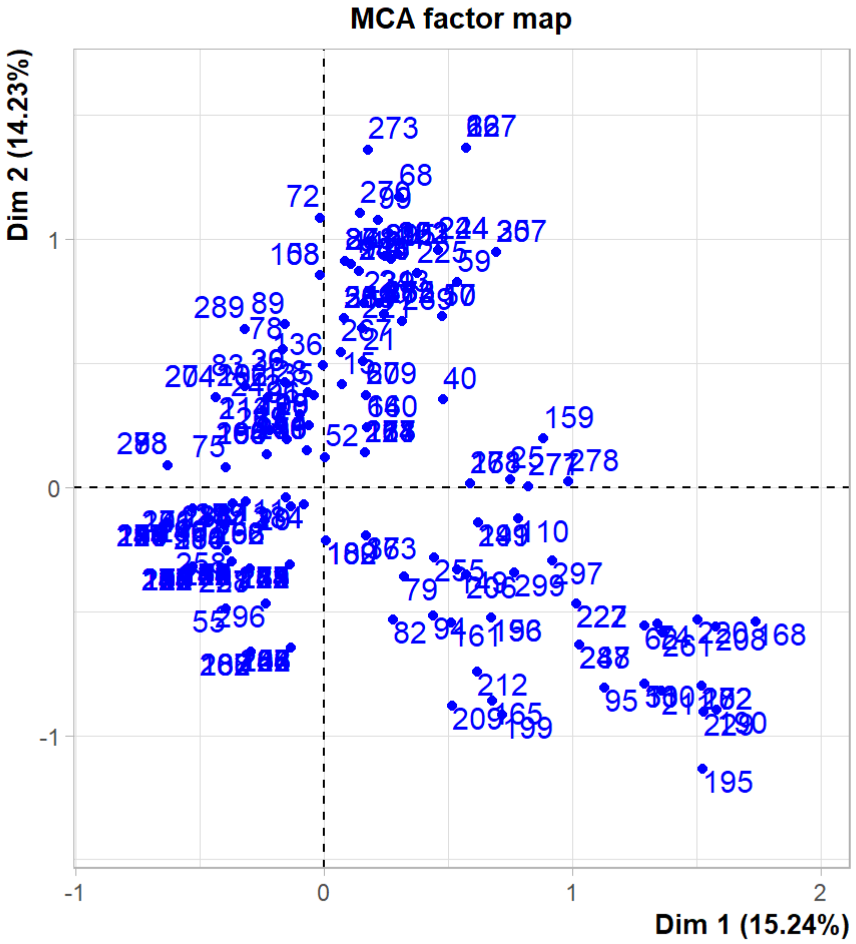
```
## 
## Call:
## MCA(X = tea_time, graph = FALSE) 
## 
## 
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               0.279   0.261   0.219   0.189   0.177   0.156   0.144
## % of var.             15.238  14.232  11.964  10.333   9.667   8.519   7.841
## Cumulative % of var.  15.238  29.471  41.435  51.768  61.434  69.953  77.794
##                        Dim.8   Dim.9  Dim.10  Dim.11
## Variance               0.141   0.117   0.087   0.062
## % of var.              7.705   6.392   4.724   3.385
## Cumulative % of var.  85.500  91.891  96.615 100.000
## 
## Individuals (the 10 first)
##                       Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## 1                    | -0.298  0.106  0.086 | -0.328  0.137  0.105 | -0.327
## 2                    | -0.237  0.067  0.036 | -0.136  0.024  0.012 | -0.695
## 3                    | -0.369  0.162  0.231 | -0.300  0.115  0.153 | -0.202
## 4                    | -0.530  0.335  0.460 | -0.318  0.129  0.166 |  0.211
## 5                    | -0.369  0.162  0.231 | -0.300  0.115  0.153 | -0.202
## 6                    | -0.369  0.162  0.231 | -0.300  0.115  0.153 | -0.202
## 7                    | -0.369  0.162  0.231 | -0.300  0.115  0.153 | -0.202
## 8                    | -0.237  0.067  0.036 | -0.136  0.024  0.012 | -0.695
## 9                    |  0.143  0.024  0.012 |  0.871  0.969  0.435 | -0.067
## 10                   |  0.476  0.271  0.140 |  0.687  0.604  0.291 | -0.650
##                         ctr   cos2  
## 1                      0.163  0.104 |
## 2                      0.735  0.314 |
## 3                      0.062  0.069 |
## 4                      0.068  0.073 |
## 5                      0.062  0.069 |
## 6                      0.062  0.069 |
## 7                      0.062  0.069 |
## 8                      0.735  0.314 |
## 9                      0.007  0.003 |
## 10                     0.643  0.261 |
## 
## Categories (the 10 first)
##                          Dim.1     ctr    cos2  v.test     Dim.2     ctr    cos2
## black                |   0.473   3.288   0.073   4.677 |   0.094   0.139   0.003
## Earl Grey            |  -0.264   2.680   0.126  -6.137 |   0.123   0.626   0.027
## green                |   0.486   1.547   0.029   2.952 |  -0.933   6.111   0.107
## alone                |  -0.018   0.012   0.001  -0.418 |  -0.262   2.841   0.127
## lemon                |   0.669   2.938   0.055   4.068 |   0.531   1.979   0.035
## milk                 |  -0.337   1.420   0.030  -3.002 |   0.272   0.990   0.020
## other                |   0.288   0.148   0.003   0.876 |   1.820   6.347   0.102
## tea bag              |  -0.608  12.499   0.483 -12.023 |  -0.351   4.459   0.161
## tea bag+unpackaged |   0.350   2.289   0.056   4.088 |   1.024  20.968   0.478
## unpackaged           |   1.958  27.432   0.523  12.499 |  -1.015   7.898   0.141
##                       v.test     Dim.3     ctr    cos2  v.test  
## black                  0.929 |  -1.081  21.888   0.382 -10.692 |
## Earl Grey              2.867 |   0.433   9.160   0.338  10.053 |
## green                 -5.669 |  -0.108   0.098   0.001  -0.659 |
## alone                 -6.164 |  -0.113   0.627   0.024  -2.655 |
## lemon                  3.226 |   1.329  14.771   0.218   8.081 |
## milk                 2.422 |   0.013   0.003   0.000   0.116 |
## other                5.534 |  -2.524  14.526   0.197  -7.676 |
## tea bag             -6.941 |  -0.065   0.183   0.006  -1.287 |
## tea bag+unpackaged  11.956 |   0.019   0.009   0.000   0.226 |
## unpackaged          -6.482 |   0.257   0.602   0.009   1.640 |
##
## Categorical variables (eta2)
##                       Dim.1 Dim.2 Dim.3  
## Tea                 | 0.126 0.108 0.410 |
## How                 | 0.076 0.190 0.394 |
## how                 | 0.708 0.522 0.010 |
## sugar               | 0.065 0.001 0.336 |
## where               | 0.702 0.681 0.055 |
## lunch               | 0.000 0.064 0.111 |
```

```
plot(mca, invisible=c("ind"), habillage = "quali")
```



The MCA is a very useful tool to analyze a non-numerical, nominal cathegorcal, qualitative data. Among others, it provides insights into existing patterns in the data. From the example of the MCA above,that explors tea-drinking habits, we can get insights into various aspects of the data at once. Also, some interesting relationships are revealed. For example, we can see that Earl Grey is more likely to be drunk with milk, than lemon or alone. It’s also more likely to be drunk with sugar than without. The black tea, on the other hand, is more likely to be taken without sugar, as well as more likely to be enjoyed with lemon than with milk. Most individuals don’t drink tea with lunch.

```
plot(mca, invisible=c("var"), habillage = "quali")
```



#

The MCA above shows exclusivly the distribution of data concerning individuals, suggesting similarities and dissimilarities among them.

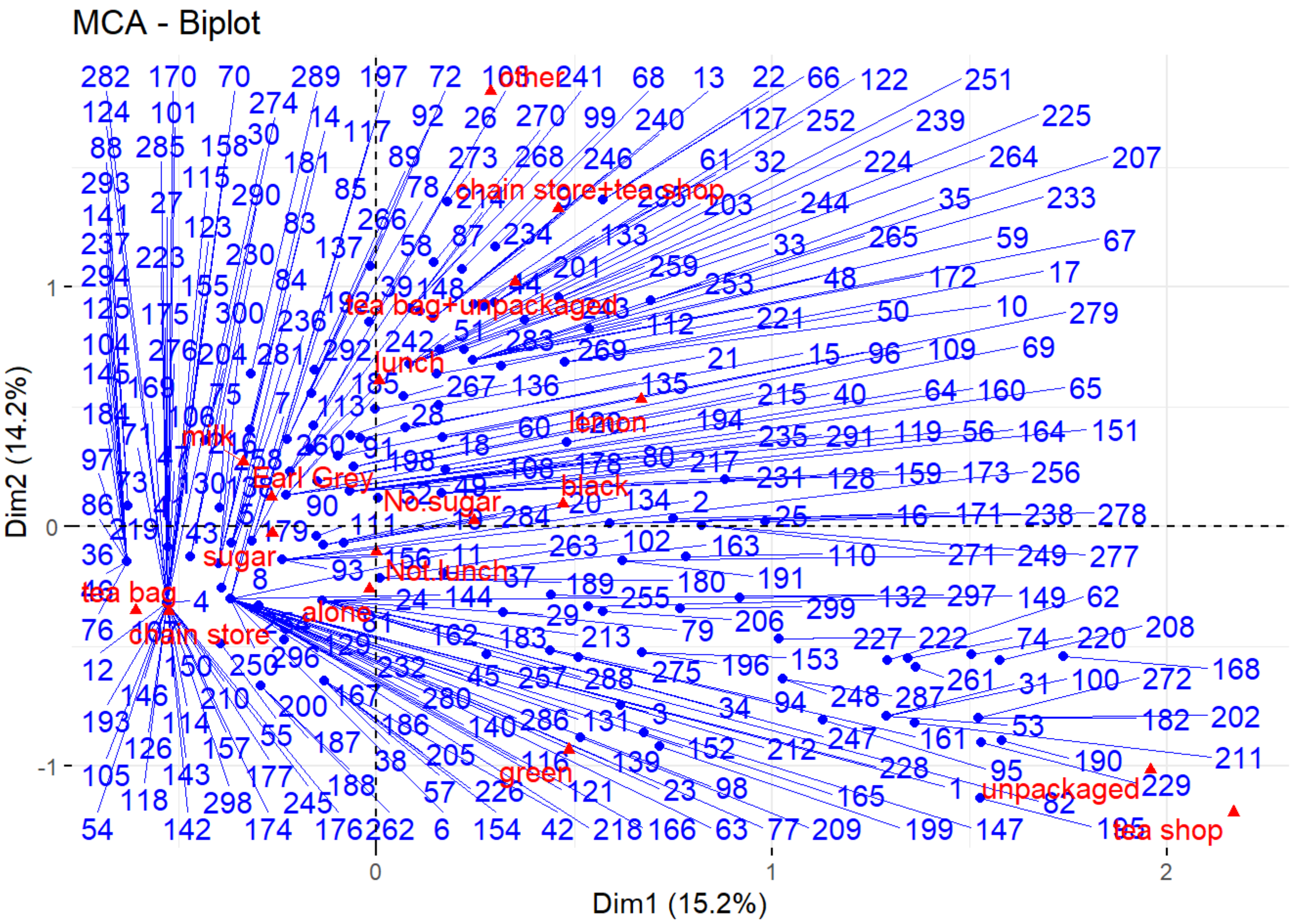
```
library(factoextra)
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFC
Z
```

```
data("tea_time")
```

```
## Warning in data("tea_time"): data set 'tea_time' not found
```

```
res.mca <- MCA(tea_time, graph=FALSE)
fviz_mca_biplot(res.mca, repel = TRUE, ggtheme = theme_minimal())
```



This overwhelming Biplot shows both, variables and individuals at the same time, highliting relationships among them. Again, the distance measures the similarity and dissimilarity among the variables and individuals.