Assignment 2: MCA and FAMD

Teacher: Miguel Reyes

I – Multiple correspondence analysis

A – Dataset

The dataset we will be using is downloaded from the package FactoMineR and is named “Tea”. It is composed of a survey asked to 300 people (the data set has 300 rows), with 36 different questions (36 different variables, which means 36 columns) about their tea consumption, how they perceive the product, and personal questions.

B – Why should we use MCA

Multiple correspondence analysis (MCA) is a dimensional reduction method used for categorical values and not continuous data. When we used CA for the previous assignment there was only two variables, here the data set has more than 2 categorical variables, which explains the use of MCA. In this dataset we will keep categorical variables and get rid of the continuous values (like the age).

What are the questions we want to ask ourselves threw this project?

Using MCA, we want to highlight the link between some of the variable of this dataset, and to see the links between some of the values of the variables to determine correlation between the values and visualize the data set in a more comprehensible way. For this example, we will select the following variables of the dataset: home (if they drink tea at home), evening (if they drink tea during the evening), sugar (if they use sugar or not), breakfast and lunch.

The questions we can ask are if there is a link between the time of the day where people drink tea and their usage of sugar, or a link between the time are which they drink tea and if they are at home, or if they use sugar or not when they are at home or outside.

C – MCA explained

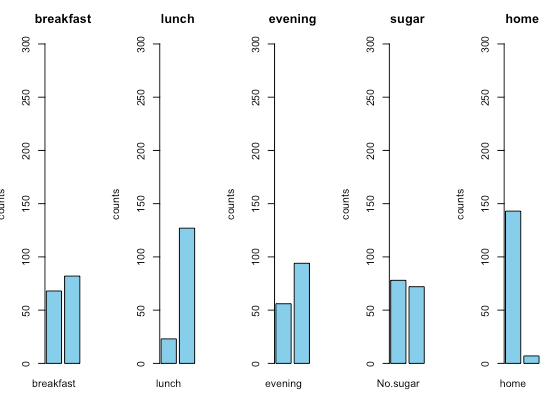
Multiple correspondence analysis (MCA) is an extension of correspondence analysis (CA) that we used for the first assignment to analyze data using a contingency table. MCA is an extension of CA, used to analyze more than one variable. To apply MCA, we can fist consider the indicator Matrix, which is a special representation of the basic contingency tables we used for CA. However, the computation for MCA is not made using an indicator matrix, that could be very large if there are many cases for the variables.

To apply the MCA, we will consider the Burt Matrix, which is the inner product of the indicator matrix. The Burt Matrix will have different cross tabulations appearing, depending on the number of variables considered. In the case of two variables, there will be 4 cross tabulations: variable 1 with variable 1, variable 1 with variable 2, variable 2 with variable 2 and finally variable 2 with variable 1. The cross tabulations of a variable against itself should be a symmetrical and diagonal matrix, and the partition should add up to the number of observations.

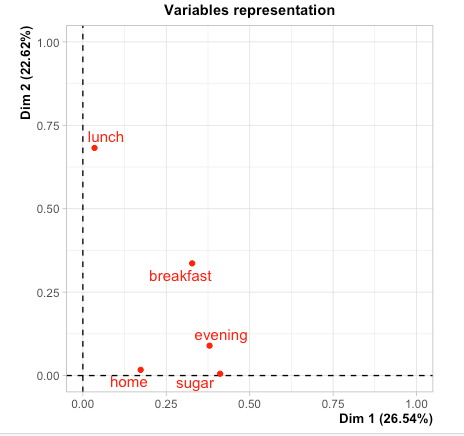
Now that we created a Burt Matrix, we can get the cross tabulation of variables by pair. We can now apply the CA method we previously saw.

D – Answering the questions

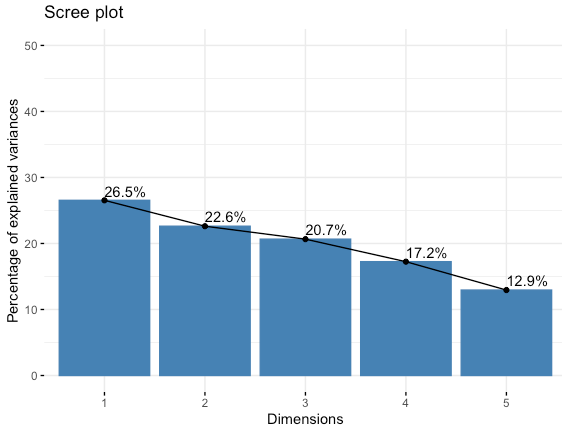
We start by importing the tea dataset, and then we select the variables we described above. We can get a first view of the distribution of those variables:



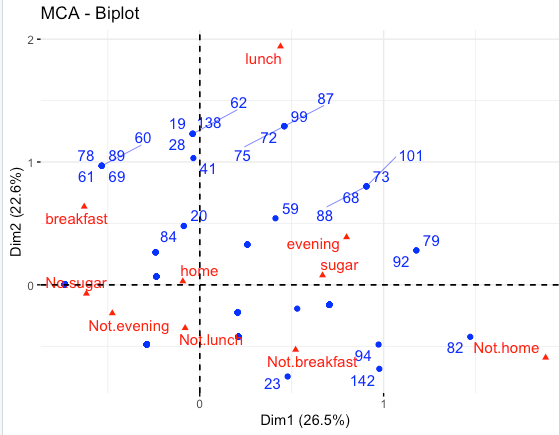
Then we can apply the MCA and get a first idea of the representation of the 5 variables we selected in a 2-dimensional space. We can also print the eigenvalues, in order to see how well the variables are represented in the dimensions.



We can see that the first two dimensions represent 49% of the variance of the dataset which is considered good enough.

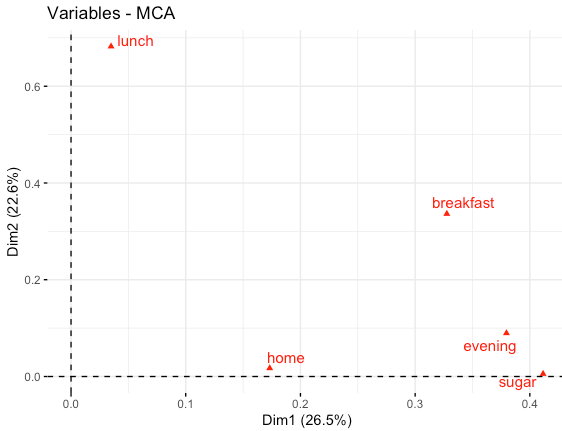


We can then make a simple 2D biplot that will be showing the correlation between different variables, and different individuals (rows):

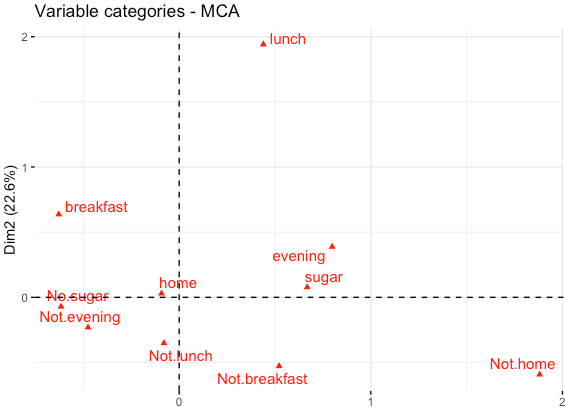


The distance represents how similar rows are one to the other, and the same applies for the columns. We can see here that rows 99 and 87 have a similar profile, and that the variables evening, and sugar are correlated.

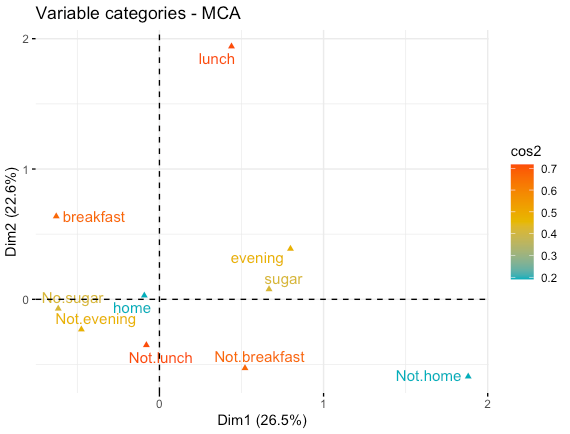
We can then make a plot for the variables, so see on which dimension which variables are correlated. Here we can see that evening and sugar are well correlated over the first dimension as an example.



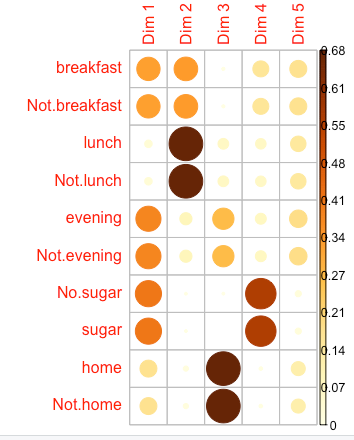
We can then visualize the different variables categories, as an example for lunch these are lunch and not lunch (if the people that took the survey drinks tea at that moment). The variables which are in opposite quartiles of the graph are negatively correlated. As we could expect, evening and not evening are negatively correlated as an example, but we have other interesting observations to make such as the fact that sugar and not lunch are negatively correlated. This logical since we saw before on the first plot that sugar and lunch are correlated.



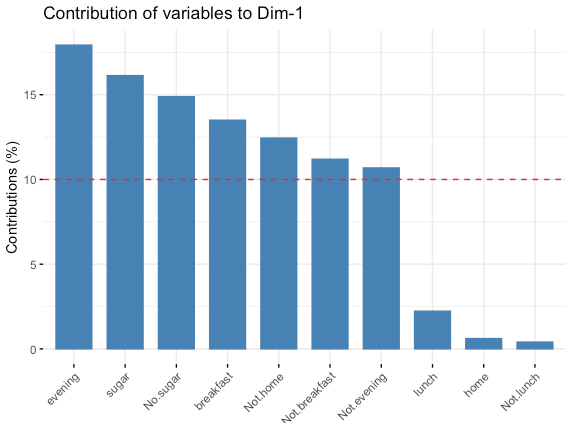
Since we know that only 49% of the variance of the dataset is represented in two dimensions, we can get the cos2 value and then make the plot again representing it: this will give us an idea of how well each variable is represented the 2D space. Here we can see that lunch and not lunch are well represented, but home and not home are not.

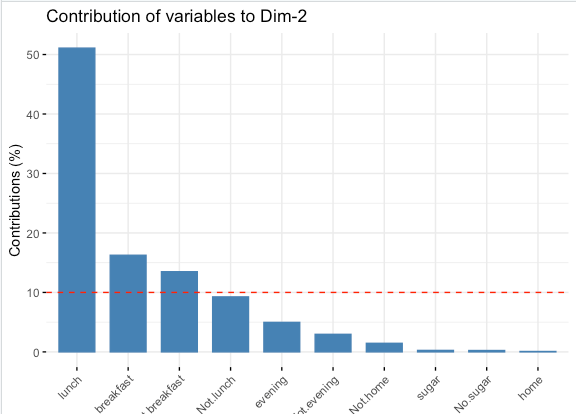


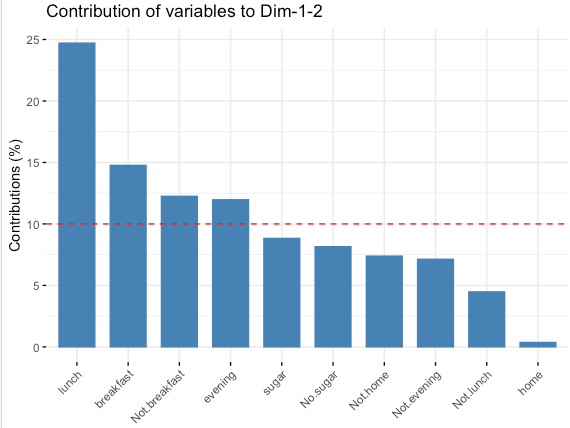
We can also make this correlation plot that will show to which dimensions the variables are correlated. Here we can see that home and not home are mostly correlated to the 3rd dimension and a little to the others, which explains why it is poorly represented in the 2D space.



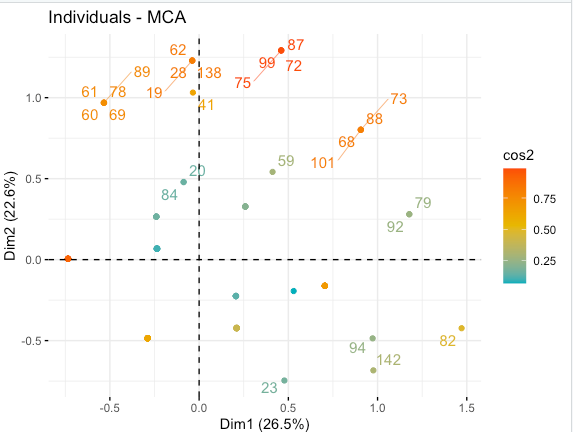
We can then make a few plots showing which variables contribute the most to the first two dimensions. The red line shows the average expected contribution of a variable to a dimension. We can see that evening and sugar contribute the most to dimension 1, and lunch and breakfast to the 2nd dimension. Over the 2 dimensions, lunch and breakfast contribute the most. Home is the least contributing variable, which we could have expected since we saw above it is mostly represented in dimension 3.



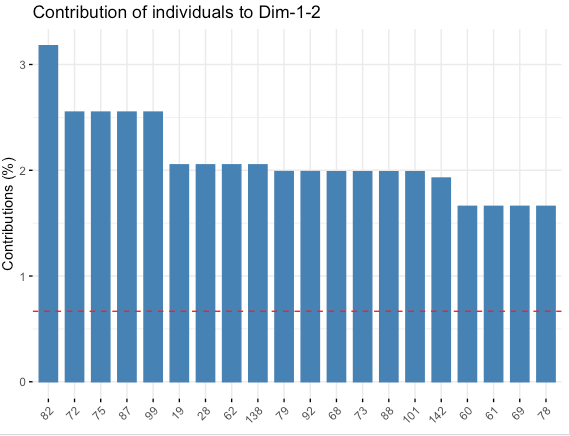




Now that we saw the representation of variables, we will see the representation of individuals. We retrieve the cos2, contribution and coordinates of individuals (rows), and we can then display some graphs. Using the cos2 value on this 2-dimensional graph, we can see that some individuals such as rows 99 or 87 are very well represented on these 2 dimensions, but rows 20 and 23 are not.

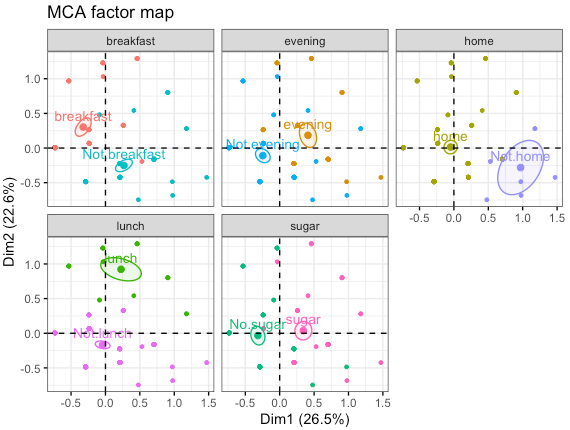


We can now display the contribution of individuals over the 2 dimensions. We can see some individuals such as rows 82 and 72 contribute the most, however this isn’t as representative as for the columns since we have 150 rows.



We can now try to answer our questions, plotting different MCA factor maps of variables. This will allow us to see the links between different variables. We wanted to determine the correlation between the variables we selected. When looking at the graphs we displayed above, we can answer those question with a certain degree of certainty (since we do not represent the global variance of the data set with those 2 dimensions).

To make those relations clearer, we use this final plot that shows the repartition of data based on the variables and the confidence interval.



We can see that as expected, breakfast and home seem correlated, and not breakfast and not home as well (most people take their breakfast at home). We can also see that sugar, evening and lunch are in the same quartile which means that people tend to use sugar in their tea at these parts of the day. Lunch, evening and home/not home do not have a correlation as clear as breakfast, since some people will be home and other will not. Moreover, breakfast and sugar/not sugar don’t have a demarcation as clear as the other two meals on the two dimensions, because some people will use sugar and other will not.