Assignment 2: MCA and FAMD

Teacher: Miguel Reyes

Barbe Victor 403715

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, which we will summarize quickly.

The principle of CA is then to compute the residual of each sell in the table, which is the difference between the value we would observe is the variables were uncorrelated, and the value we actually observe (here the table will be a part of the Burt Matrix). The bigger the residual is (negative or positive), the bigger the correlation is. We can then also compute the row and column mass, which are the added values of that column or row. Then to get the expected value of a cell, we just multiply the row mass by the column mass.

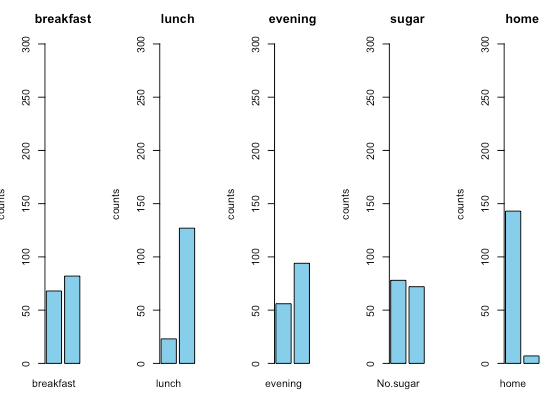
Now that we have the expected and the actual value of each cell, we can compute the table of residuals. This table is going to give us more information about the data. For our example, if we were to get a negative value in the residual table two attributes, this means that they are less likely to be linked.

Now to get more accurate information, we can compute the indexed Residual. This will compensate for the difference in the sample size of attributes (if someone did more activities as an example). To do so, we divide the residual value we calculated and divide it by the expected value.

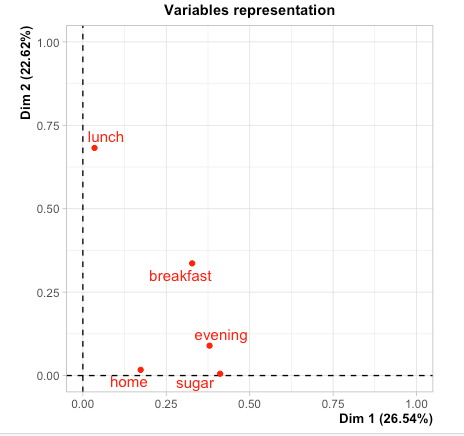
Using this method, we now get a percentage relation between attributes. Then to compute the CA, we take the SVD of this index residual matrix. The result of SVD we get will give us a vector, and two matrices that will give us the information needed about the dimensions, columns and rows to get the CA. Using the left and right singular vector, we can now compute the CA and plot the result.

D – Answering the questions

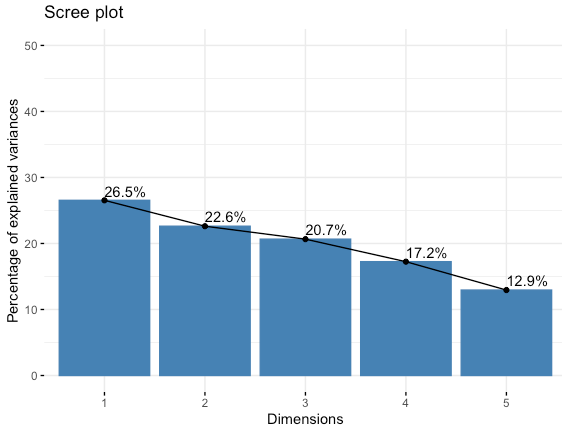
We start by importing the tea dataset, and then we select the variables we described above. We can get a first view of the distribution of those variables:



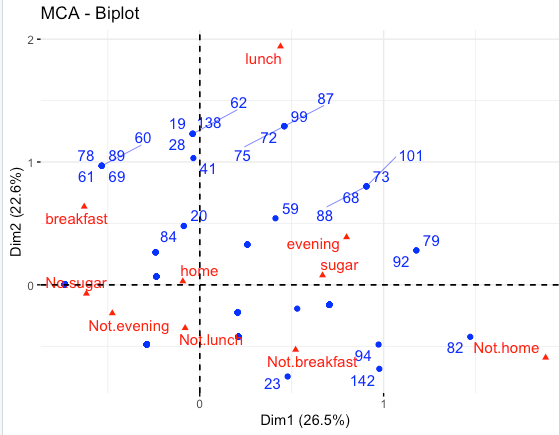
Then we can apply the MCA and get a first idea of the representation of the 5 variables we selected in a 2-dimensional space. We can also print the eigenvalues, in order to see how well the variables are represented in the dimensions.



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

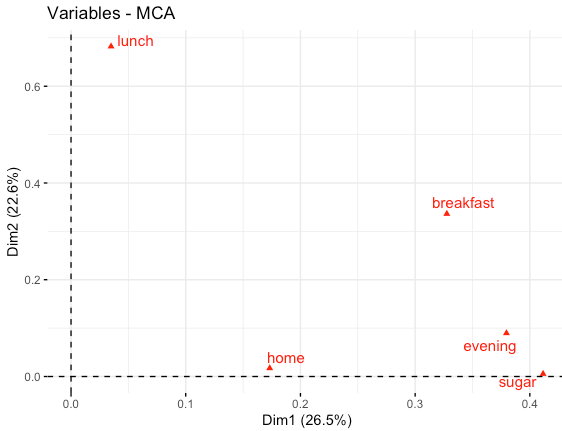


We can then make a simple 2D biplot that will be showing the correlation between different variables, and different individuals (rows):

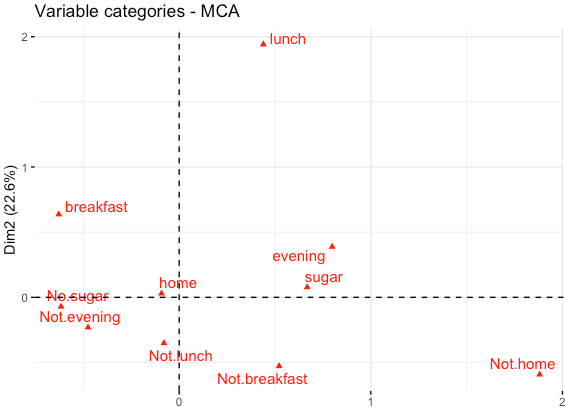


The distance represents how similar rows are one to the other, and the same applies for the columns. We can see here that rows 99 and 87 have a similar profile, and that the variables evening, and sugar are correlated.

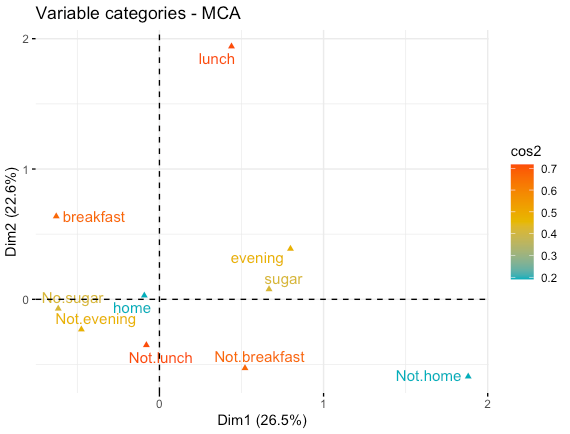
We can then make a plot for the variables, so see on which dimension which variables are correlated. Here we can see that evening and sugar are well correlated over the first dimension as an example.



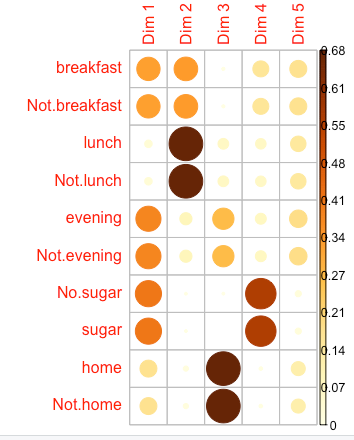
We can then visualize the different variables categories, as an example for lunch these are lunch and not lunch (if the people that took the survey drinks tea at that moment). The variables which are in opposite quartiles of the graph are negatively correlated. As we could expect, evening and not evening 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.



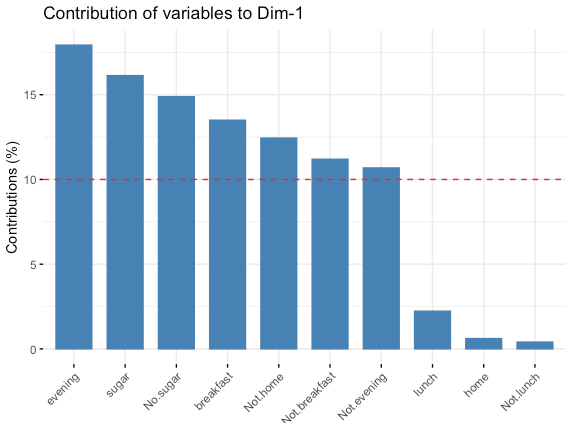
Since we know that only 49% of the variance of the dataset is represented in two dimensions, we can get the cos2 value and then make the plot again representing it: this will give us an idea of how well each variable is represented the 2D space. Here we can see that lunch and not lunch are well represented,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

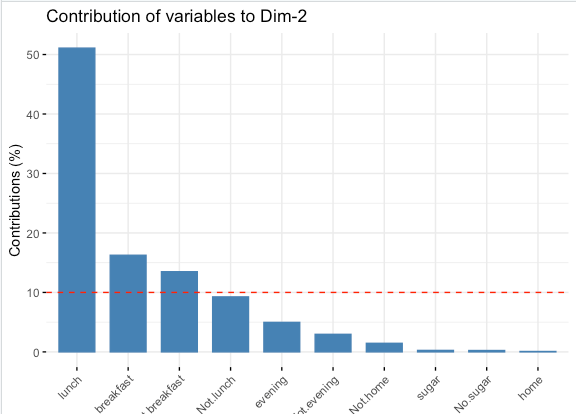


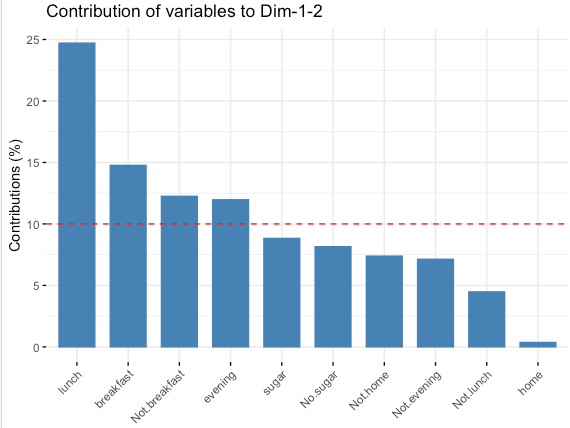
We can also make this correlation plot that will show to which dimensions the variables are correlated. Here we can see that home and not home are mostly correlated to the 3rd dimension and a little to the others, which explains why it is poorly represented in the 2D space.



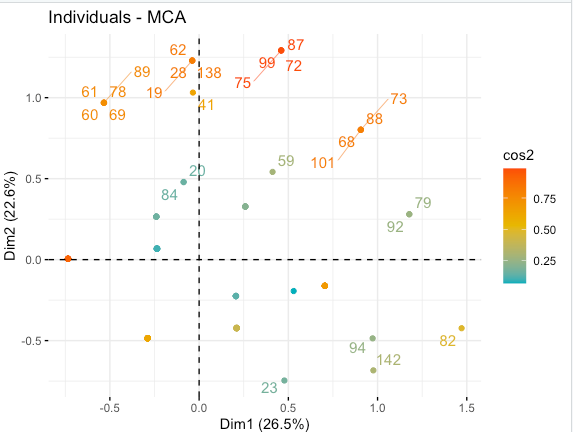
We can then make a few plots showing which variables contribute the most to the first two dimensions. The red line shows the average expected contribution of a variable to a dimension. We can see that evening and sugar contribute the most to dimension 1, and lunch and breakfast to the 2nd dimension. 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.



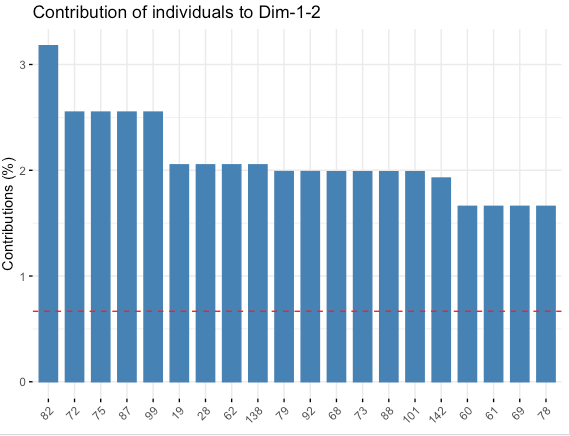




Now that we saw the representation of variables, we will see the representation of individuals. We retrieve the cos2, contribution and coordinates of individuals (rows), and we can then display some graphs. Using the cos2 value on this 2-dimensional graph, we can see that some individuals such as rows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

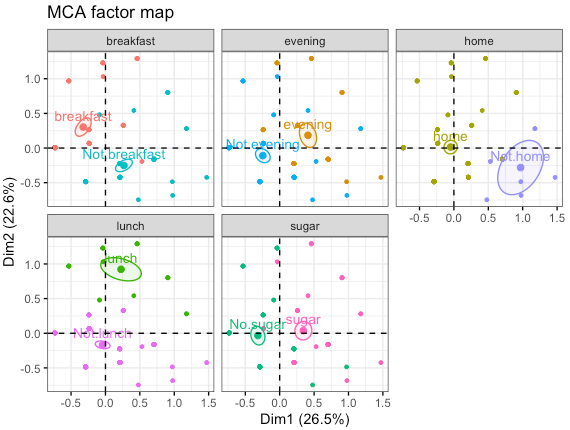


We can now display the contribution of individuals over the 2 dimensions. We can see some individuals such as rows 82 and 72 contribute the most, however this isn’t as representative as for the columns since we have 150 rows.



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.

II – Factor analysis of mixed data

A – Dataset

The dataset we will be using was found on Kaggle. It is composed of Exam results of students studying in a school. The data set is composed of the results of an exam in 3 different classes (math, reading and writing) as well as other information on the student such as the level of education of the parents, their gender, ethnicity, the preparation they add for the exam and the type of lunch they could afford.

B – Why should we use FAMD

We need to use FAMD for this part of the project because we have mixed types of data (both continuous and categorical variables. In this case, we cannot use PCA since it only works on continuous variables, and we cannot use CA or MCA since it only works on categorical variables. Since our dataset is mixed, we will have to use FAMD to get results.

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

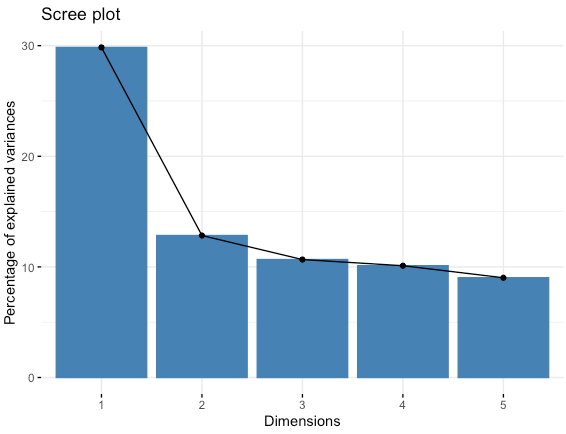
In this project, we would like to understand whether there is a link between characteristics of a student’s life and how well they perform at school.

C – FAMD explained

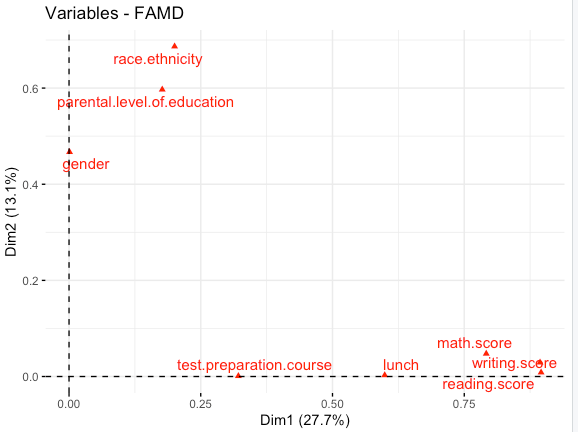
Factor Analysis of Mixed Data (FAMD) is a factorial method which is used on data tables which present both quantitative and qualitative variables. To treat the different type of data, FAMD works as PCA for quantitative variables and as MCA for qualitative variables. The methodology of MCA is explained in the other part of this report, and we will summarize again the explanation of PCA.

D – Answering the question

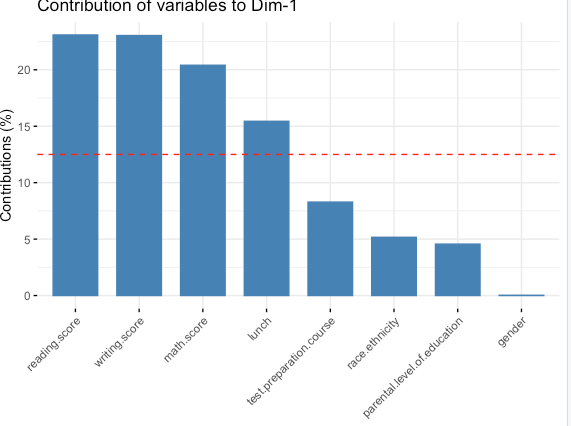
First of all, we upload the dataset in our code and use the ‘as.factor’ method to transform the categorical values. Here, it will be used for ethnicity, gender, lunch and preparation. We can then print the eigenvalues and display a graph to see how much each dimension describes the data.

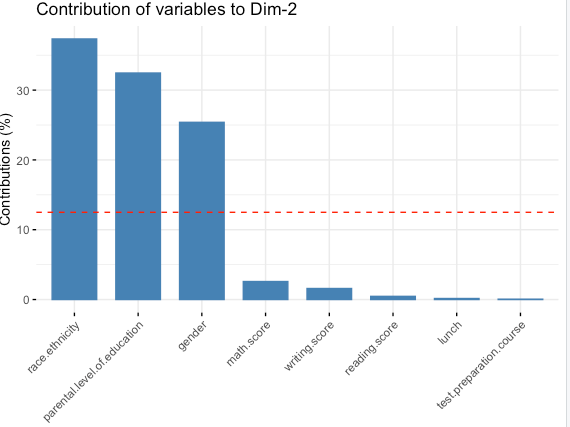


Here we get that the first two dimensions are describing 42.7% of the data set which isn’t ideal but will be enough to observe correlations and links. Then we can plot the following graph that allows to see the contribution of each variable to the first two dimensions. We can see that parental level of education and ethnicity contribute a lot to the 2nd dimension, while the different score contributes a lot to the first one.

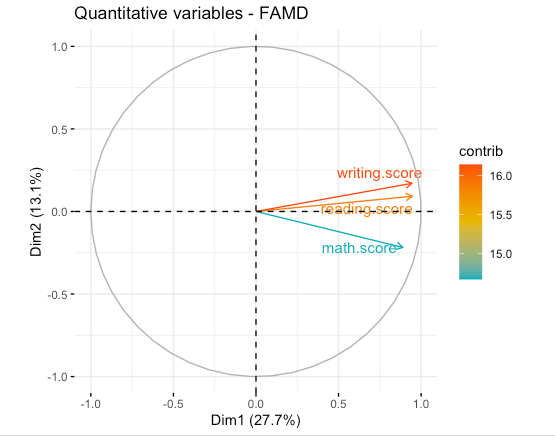


This can be observed more easily when looking at the following graphs of the contribution of each variable to each dimension. As we described above reading\_score and writing\_score contribute the most to dim1 and parental level of education and ethnicity to dimension 2.

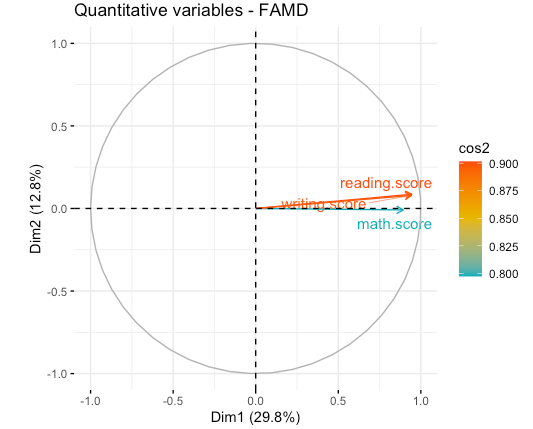




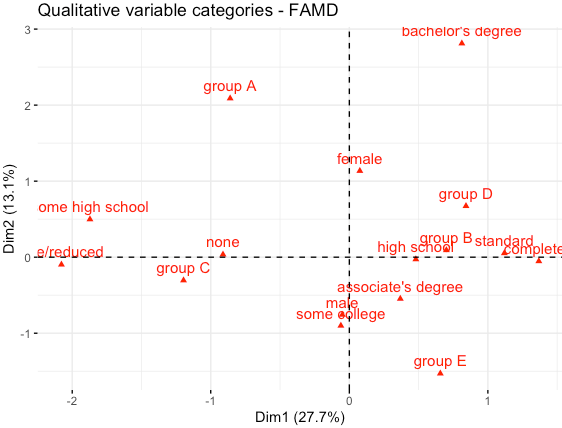
We can now visualize the quantitative variables (the scores) contribution to the dimensions. We can see that reading and writing score have the highest contributions to the first 2 dimensions. We also notice that they are all closely related, which makes sense since they all measure an aspect of the scholar level of a student.



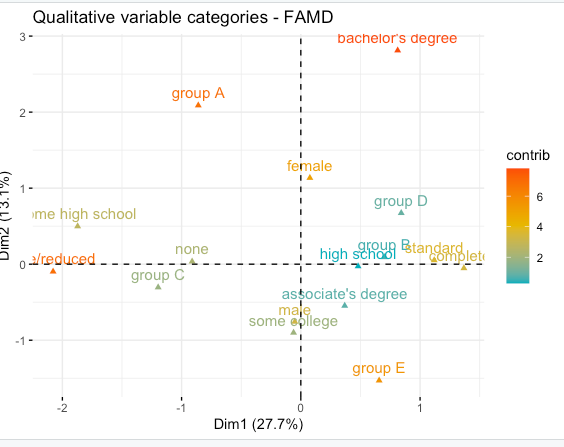
We can also get the cos2 score and plot here. Again, we can see that reading and writing score are well represented in the first two dimensions, but math score is not.

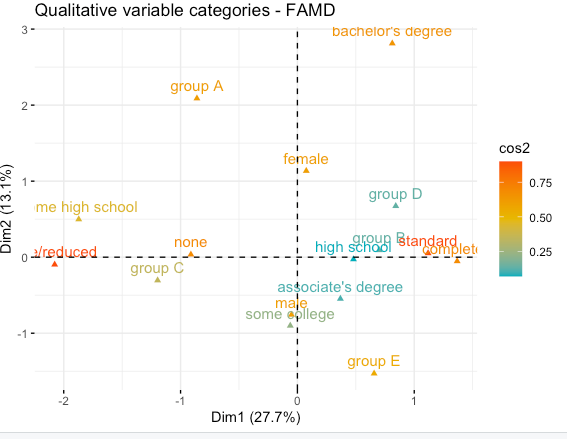


After that we can plot the qualitative variables and their different categories on the two-dimensional space. We can see that some variables are correlated and other are not. As an example, having parents that have a bachelor’s degree and being in group of ethnicity C are negatively correlated, which means belonging to those two categories at the same time is very unlikely to happen.



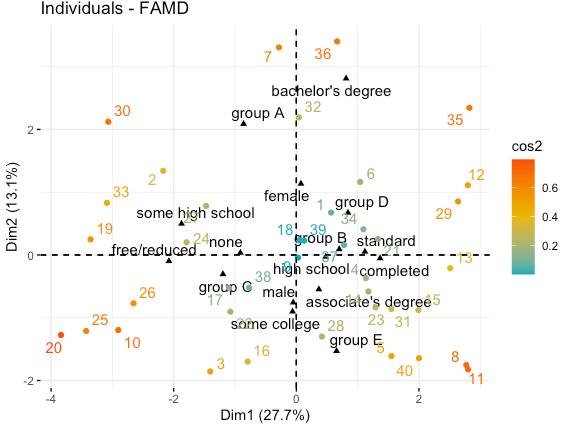
After that, we can see the contribution of the variables to the first dimensions, as well as their cos2 score. We can see that bachelor’s degree, Group A and standard are well represented in the first 2 dimensions as an example, but high school and group D are not.

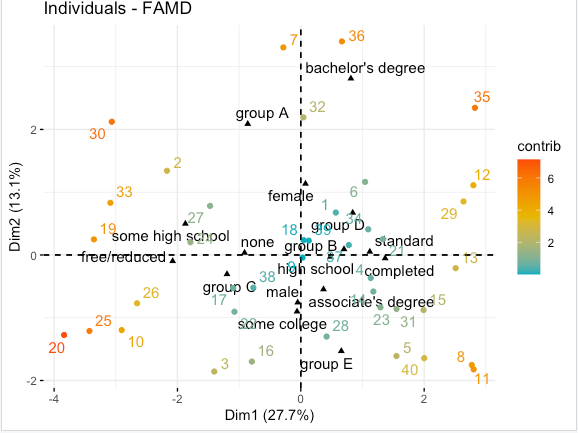




After looking at the different columns (variables), we can now look at the individuals and see their cos2 score as well as their contributions to the first 2 dimensions. We can see that some of the samples are much more accurately represented and contribute more to the two-dimensional space. These samples seem to be correlated to the variables that also were represented well in the 2D space: we can see that samples 32, 36 and 7 which are close to Group A and bachelor’s degree that where well represented are also well represented.

On the opposite, the samples close to the variables that were not very well represented are not very well represented as well, and don’t contribute a lot to the two-dimensional space.





In the end, we can see that FAMD allowed us to represent our data in a low dimensional space, while still grasping an important amount of the variance in the database. It also allowed us to discover the links between some of the variables present. We can see that the points that are closer to high level of studies such as bachelors degree have high grade levels (when taking a look at the data).

We can also see as an example that male is negatively correlated to bachelor’s degree (a high level of study), while female is corelated positively.

Female