Select Topic
Project 2 : MCA and FAMD

Gaucher Pierre-Louis 403783
Barbe Victor 403715

Part I : Multiple Correspondence Analysis

1) What is the dataset about ?

For this study, we have chosen the dataset “tea” from FactoMiner Library.
This dataset corresponds to answers of three hundred individuals over different topics related to tea beverage. Questions asked are related to three specific topics :

- The way individuals are drinking tea (place, time…)
- The way individuals are considering the product (healthy, friendly drinking…)
- The effects of the beverage on the individuals (relaxing, exciting…)

2) Why is it relevant to apply MCA ?

In a previous project we had the opportunity to implement CA in order to perform dimensional reduction, allowing us to better understand the dataset. The difference between CA and MCA is the number of categorical variables used. While CA can handle only 2, MCA can handle more. Our dataset is here referring to several categorical variables such as places (restaurant, work…), time (breakfast, lunch…). It explains why we have to use MCA instead of CA here.

We have to notice that some values of the dataset, such as age, caused trouble for the execution of MCA, because they were not categorical values. We have to ignore this kind of data in MCA implementation.

Because we have checked if MCA is applicable to the dataset, we will now clean it in order to perform the method in the best conditions.

Because of the huge amount of columns in this dataset, we have chosen 5 categorical variables to perform MCA :

- One refers to a specific time to drink tea : lunch
- The others are referring to places where to drink tea : home, work, tea room, restaurant.

We will perform MCA on those categorical variables in order to show any kind of correlation between a specific time to drink with several places.

3) MCA explanation

Part A : Basis of MCA

MCA is the application of CA while working on more than 2 categorical values.
With 2 categorical values, we can easily get a contingency table to perform CA, which will help us throughout the model.
It doesn't work on MCA. There is 2 equivalent that can be used to perform the model : Indicator Matrix and Burt Matrix

a) Indicator Matrix

An Indicator Matrix is the representation of a Contingency Table in a format where each categorical variable gets a binary value (0 or 1) for each rows. It is pretty easy to build because each rows corresponds to a specific event :

Sex		Brain Tumor	
M	F	Malignant	Benign
1	0	0	1

In this example, we can easily read our first line as : patient X is a male with a benign brain tumor.

However, this kind of matrix can be very large so it is not really used in MCA implementation.

b) Burt Matrix

This is the matrix used for our implementation of MCA, which is a shorter approach of a matrix than Indicator Matrix. However, the Indicator Matrix is used to build the Burt Matrix, which is the inner product of the Indicator Matrix.
It gives different cross tabulations of the data.
An example is given by the following, found on : tab2-2k7-586-588.gif

	Small	Medium	Large	Formal	Informal	0-5	6-11	11->
Small	30	0	0	10	20	14	13	3
Medium	0	25	0	11	14	8	9	8
Large	0	0	15	10	5	9	4	2
Formal	10	11	10	31	0	11	12	8
Informal	20	14	5	0	39	20	14	5
0-5	14	8	9	11	20	31	0	0
6-11	13	9	4	12	14	0	26	0
11->	3	8	2	8	5	0	0	13

We can notice that the cross tabulation between the same categorical variables always leads to a symmetrical matrix with values equal to 0 except on the diagonal.

We will now explain the process of MCA based on the dataset we have used and we will try to answer our question based on the model.

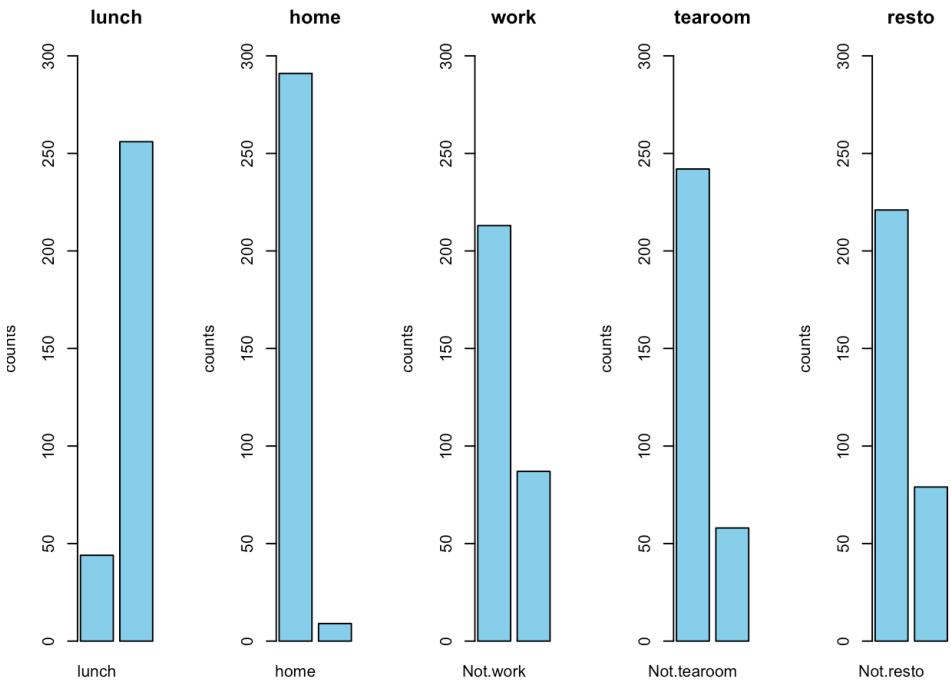
Part B : MCA functionment applied to our dataset

We have imported our dataset from FactoMiner so we first have to import this package. After checking the dataset, we have to store the categorical variable we have chosen previously using this command :

```
active<-tea[, c(4, 7, 8, 9, 11)]
```

“active” now contains the 300 rows related to columns 4, 7, 8, 9, 11 (lunch, home, work, tearoom and restaurant).

We can apply some function to check the data, such as “summary”, which will give us the repartitions inside the categorical variables, or check it more visually using graphs, such as the one above :

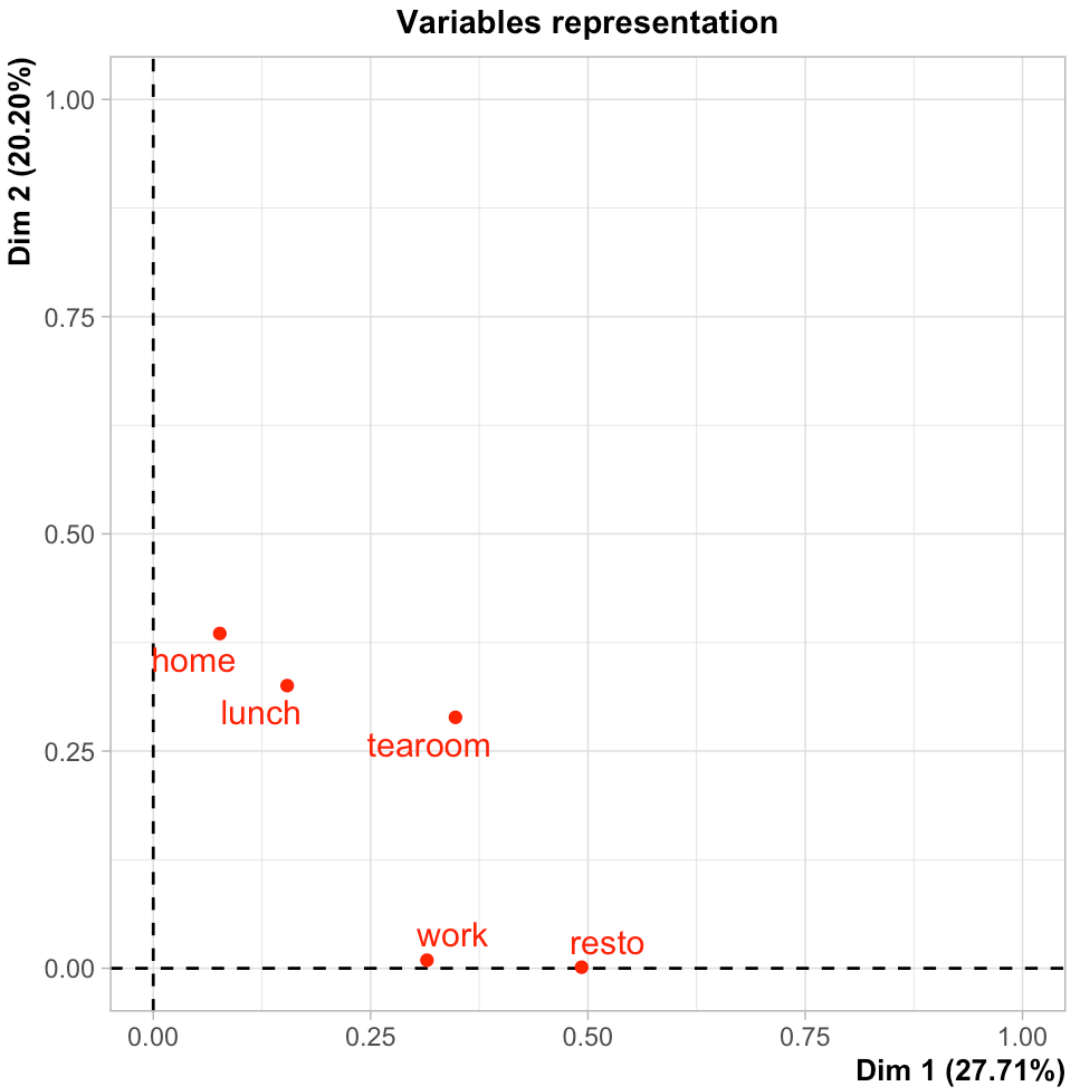


It shows us the repartition inside the categorical variable. In y axis we have the number of values while in x we have the name of the dominant variable inside the categorical variable (basically a boolean).

The repartition of our variables looks correct, we can apply MCA using :

```
ob.mca<-MCA(active, ncp = 5, graph=TRUE)
```

Where “active” corresponds to the categorical values we have kept, ncp the result dimensionality and graph gives us this representation :

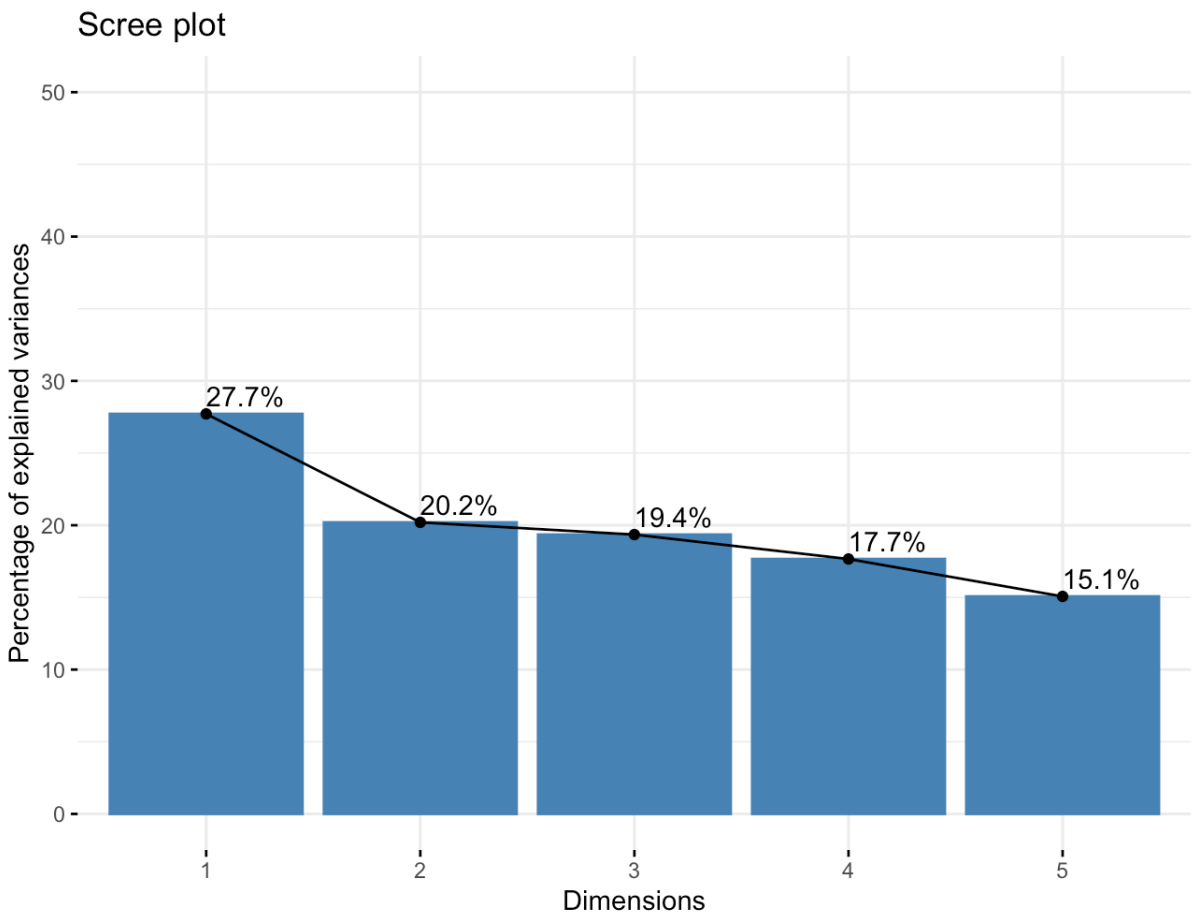


It gives us the importance of each variable in the 2 biggest dimensions found by the algorithm. We can see for example that the variable “home” impacts much more Dim2 than Dim1, while tearoom looks to impact both the same way.

The sum of the 2 biggest dimensions represent 47.91% of the total variability of the variables we have chosen.

It could be better but we managed to find variables with at least 40% variability on the 2 biggest dimensions, which was the minimum required for this study.

Further in the code, we are plotting a scree plot showing the accuracy of each dimensions :



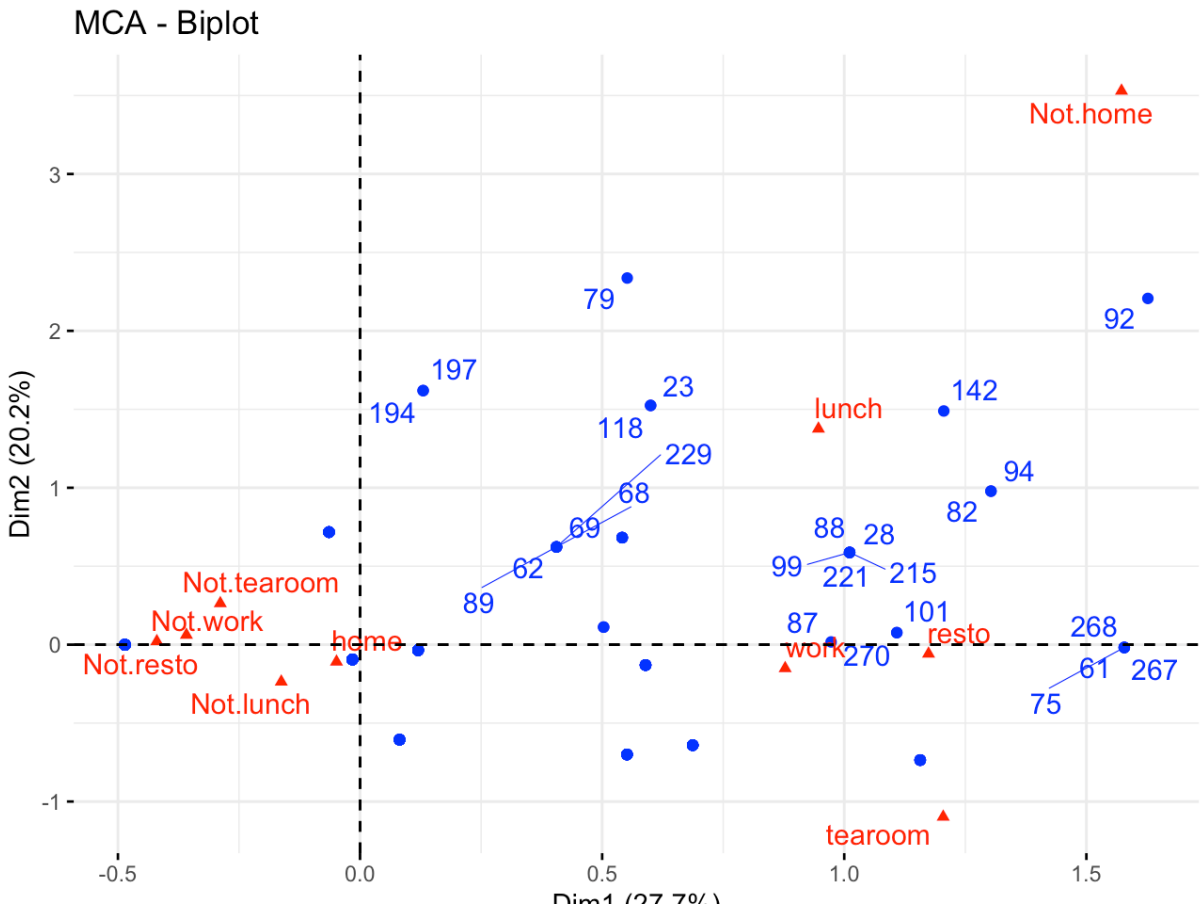
The 3 first dimensions explain almost 67% of the dataset but we will only focus on the 2 more accurate : Dim1 and Dim2.

We also got the eigen values, which gives us the same values :

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.2771431	27.71431	27.71431
Dim.2	0.2019994	20.19994	47.91426
Dim.3	0.1935427	19.35427	67.26852
Dim.4	0.1766192	17.66192	84.93044
Dim.5	0.1506956	15.06956	100.00000

The sum of our eigen values is equal to 1 which means it performs well.
Then, we will perform a biplot which will represent the data using :

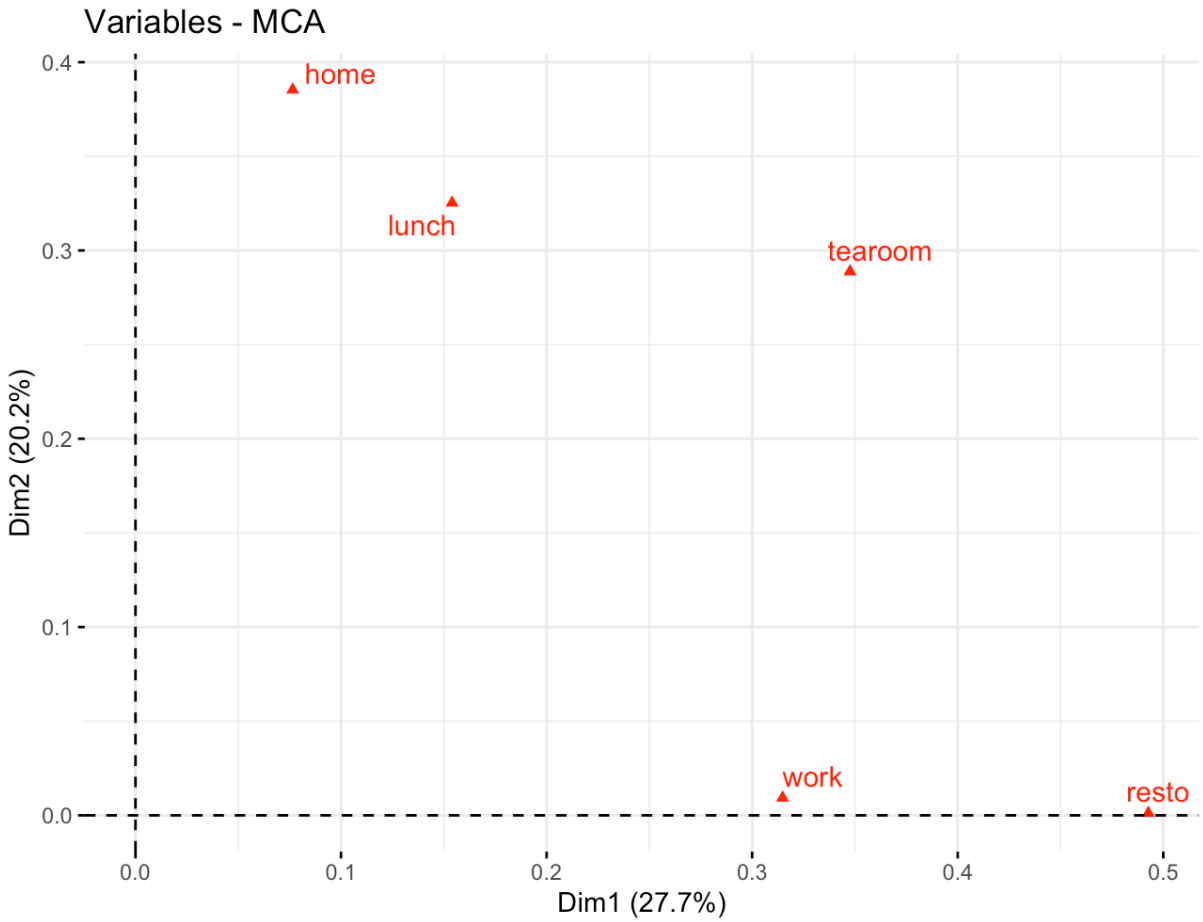
```
fviz_mca_biplot(ob.mca,
                repel=TRUE# Avoid text overlapping (can get slow if many points)
)
```



The further a point is on the X axis the more important he is for this dimension, same process with Y.

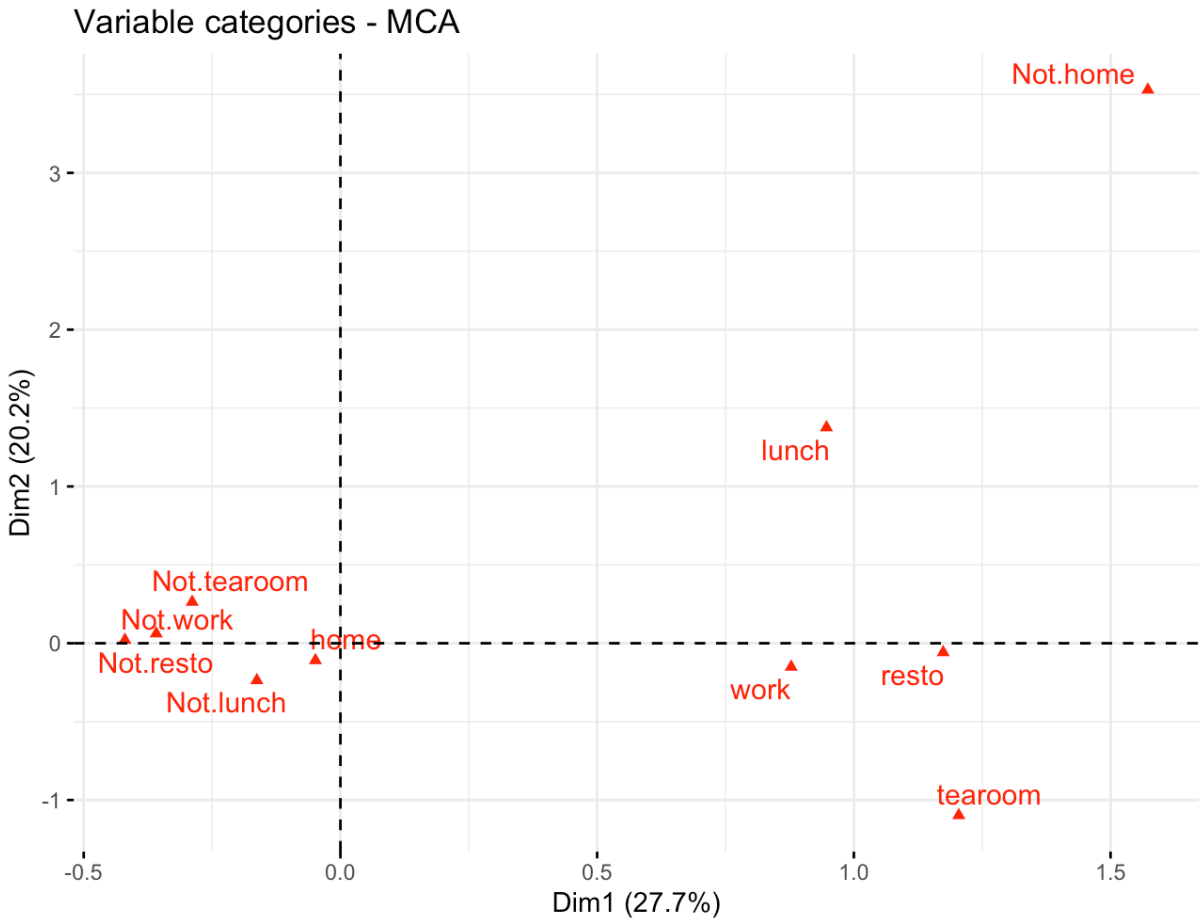
The individuals (the rows from our dataset) are represented by blue points. Because we have 300 rows it is not possible to print them all, however we can notice that the more 2 individuals are close to each other, the more similar they are. For example, rows 99, 88, 28, 215, 221 (in coordinates (1, 0.5)) are highly similar while 194 and 75 ((0, 1.6), (1.5, 0)) don't have the same relation.

It works the same for the red triangles, that corresponds to the variables of our categorical variables. We can tell that



This graph shows the influence of each variable over Dim1 and Dim2. It is quite similar to one we got in previous graphs but is more accurate because it has been scaled. It shows us that tearoom, work and resto influence a lot Dim1 while home, lunch and tearoom are more influencing Dim2.

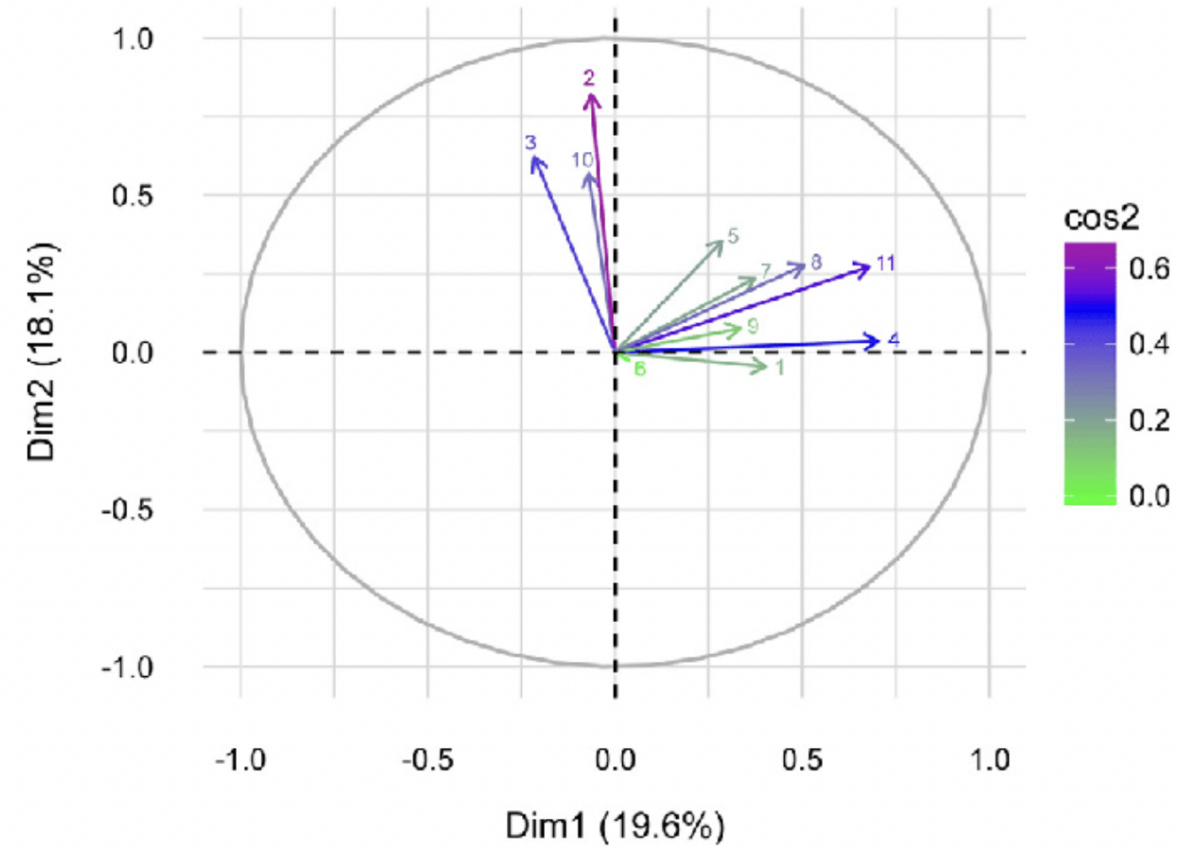
To be more precise on how each categorical variables are influencing the dimensions, we have done the same for each variable of the categorical variables of our dataset :



This graph gives us plenty of information on how variables are influencing dimensions and how they are correlated between each other.

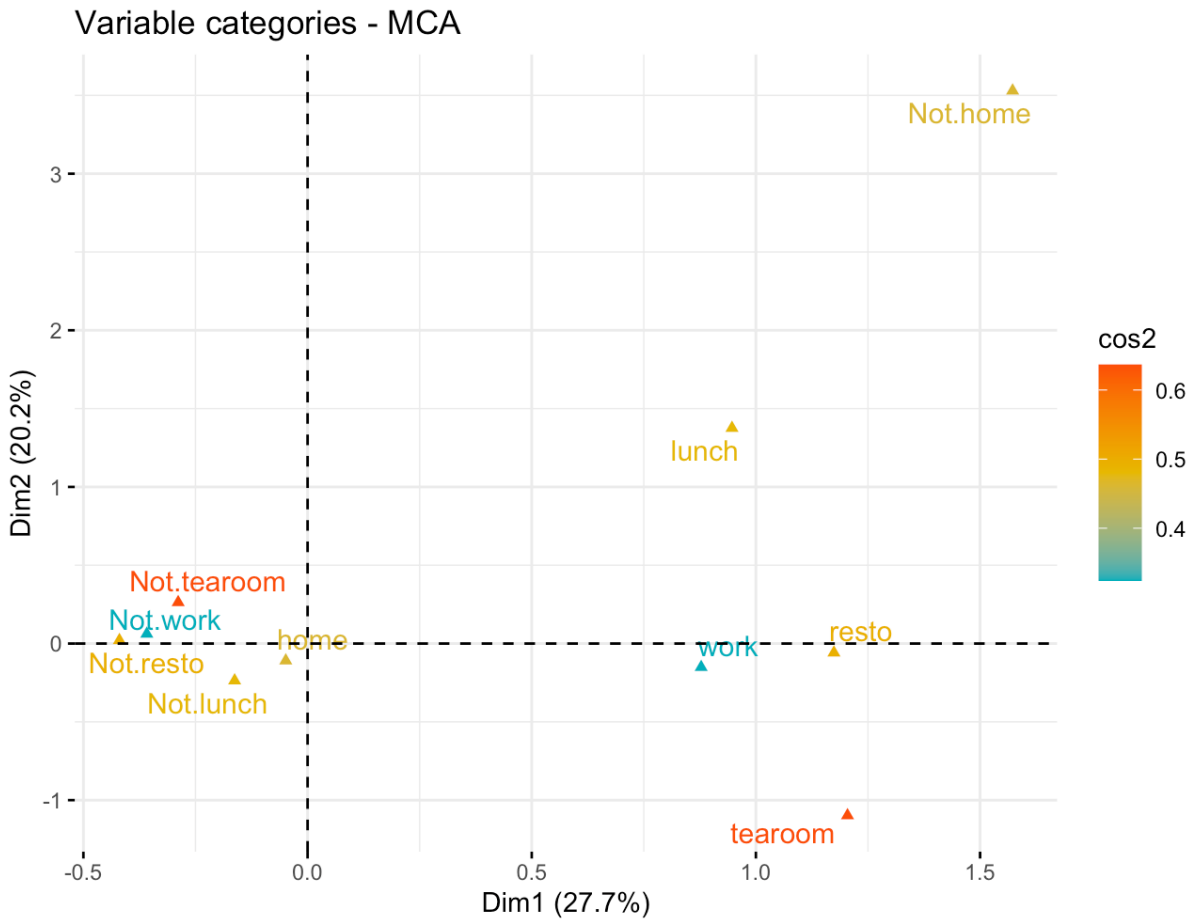
a) Two variables on opposite quadrants are negatively correlated such as work/not.work, home/not.home, tearoom/not.tearoom… It works on opposite variables obviously but it also gives information of the relationship between other variables : not lunch and not home are negatively correlated. Which looks quite obvious when we have plotted the previous graph : lunch and home were correlated (close to each other).
b) Similar variables are grouped together in the same quadrant, such as not.tearoom and not.home.
c) Distance between points and origin define the quality of their representation on the factor map. The factor map is a tool used to better understand clusters formed by correlated variables.
d) The more a category point is far from the origin, the better he is represented on the factor map.

To better understand how factor maps and cos2 are working, we will use this one found on this website.

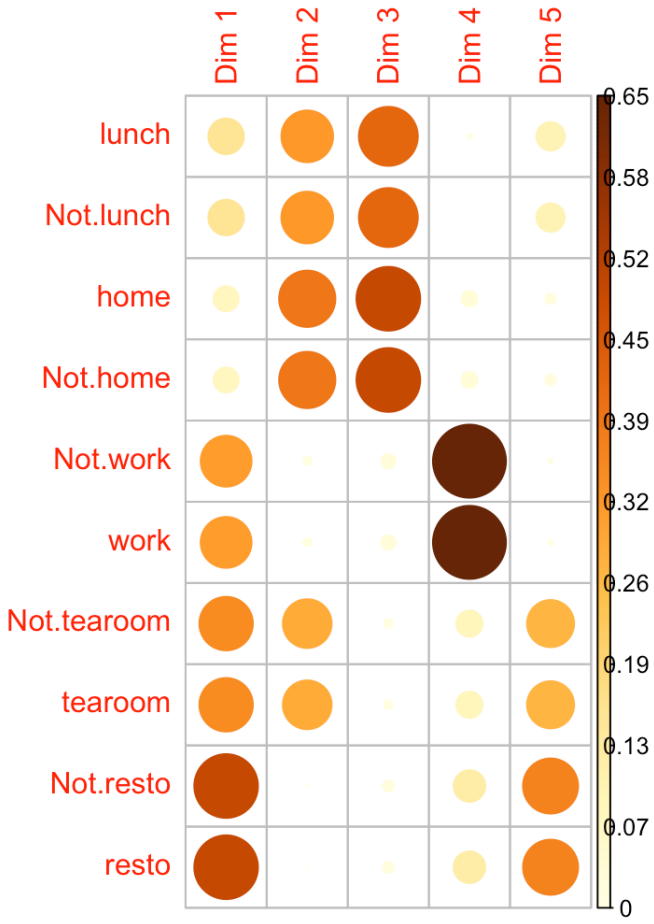


Cos2 gives an idea on how well variables are represented on a factor map, which means how they influence the dimensions. We can see that arrow 2 is the biggest of the map. This variable is associated as a big influence on Dim2 while arrow 11 has more on Dim1. Arrow 8 looks less relevant than the others with a cos almost null.

Now, we can plot the cos2 values associated with each of our variables. Not.tearoom and tearoom are the variables with the highest cos2 score. Based on the map and definitions before, it means that they influence a lot on the dimensions. It is a proof of what we have seen before when we said that tearoom categorical variable is influencing both Dim1 and Dim2. Same with work and not.work, they have the lowest cos2 values and if we get to the graph above, we have mentioned that categorical value work is influencing only Dim1, but at a lower point than tearoom and resto.

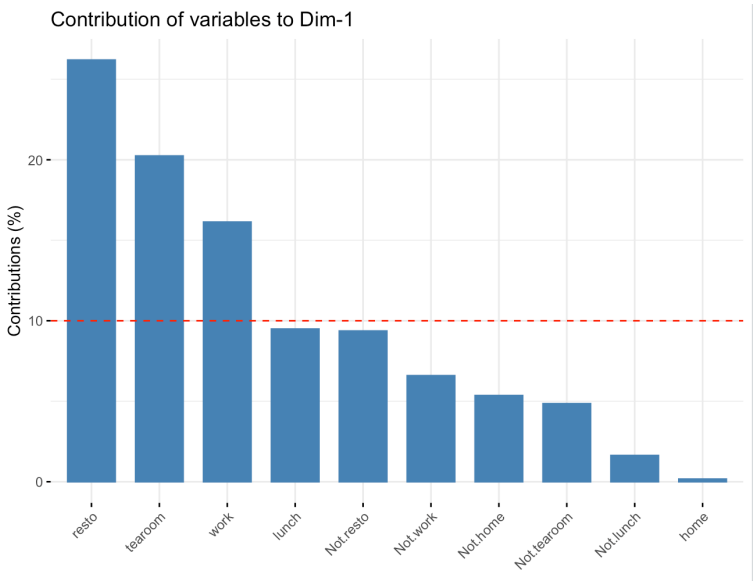

Variable categories - MCA
Not.home
lunch
Not.tearoom
Not.work
home
Not.resto
Not.lunch
work
resto
tearoom
cos2
0.6
0.5
0.4
Dim2 (20.2%)
Dim1 (27.7%)
3
2
1
0
-1
-0.5
0.0
0.5
1.0
1.5

These plots are only showing the correlation of our categorical variables to Dim1 and Dim2. What about the other ones ? Maybe the most influential variables of Dim1 and Dim2 are totally “irrelevant” for the other dimensions. We will use a corrplot to do so :

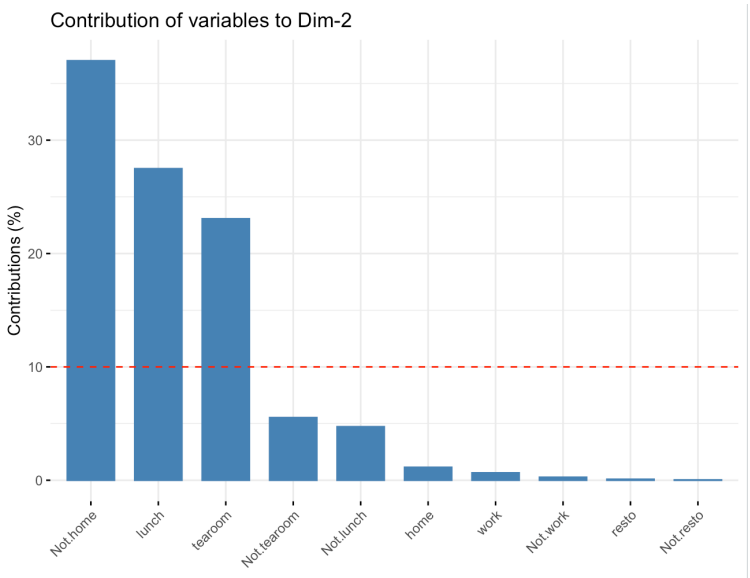


Here we are, resto is the most influential on Dim1. But we said previously that tearoom was the most influential ? It can be explained by the fact that tearoom is present both in Dim1 and Dim2 while resto is only on Dim1, even if it’s with a higher scale.
Let’s take a look to work. It had the lowest cos2 on Dim1 and Dim2 however we can see that he is highly influential on Dim4, which is not used in our study.

To get a better overview on how each variables are influencing Dim1 and Dim2, we will make plot showing their contribution percentage :

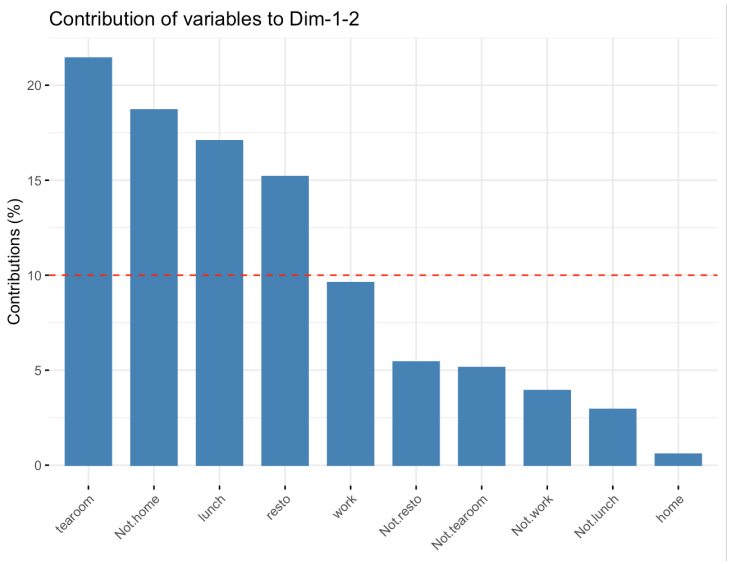


As we have seen in the corrplot above, resto is the one that most influences Dim1, followed by tearoom and work.
Let's make the same for Dim2 :

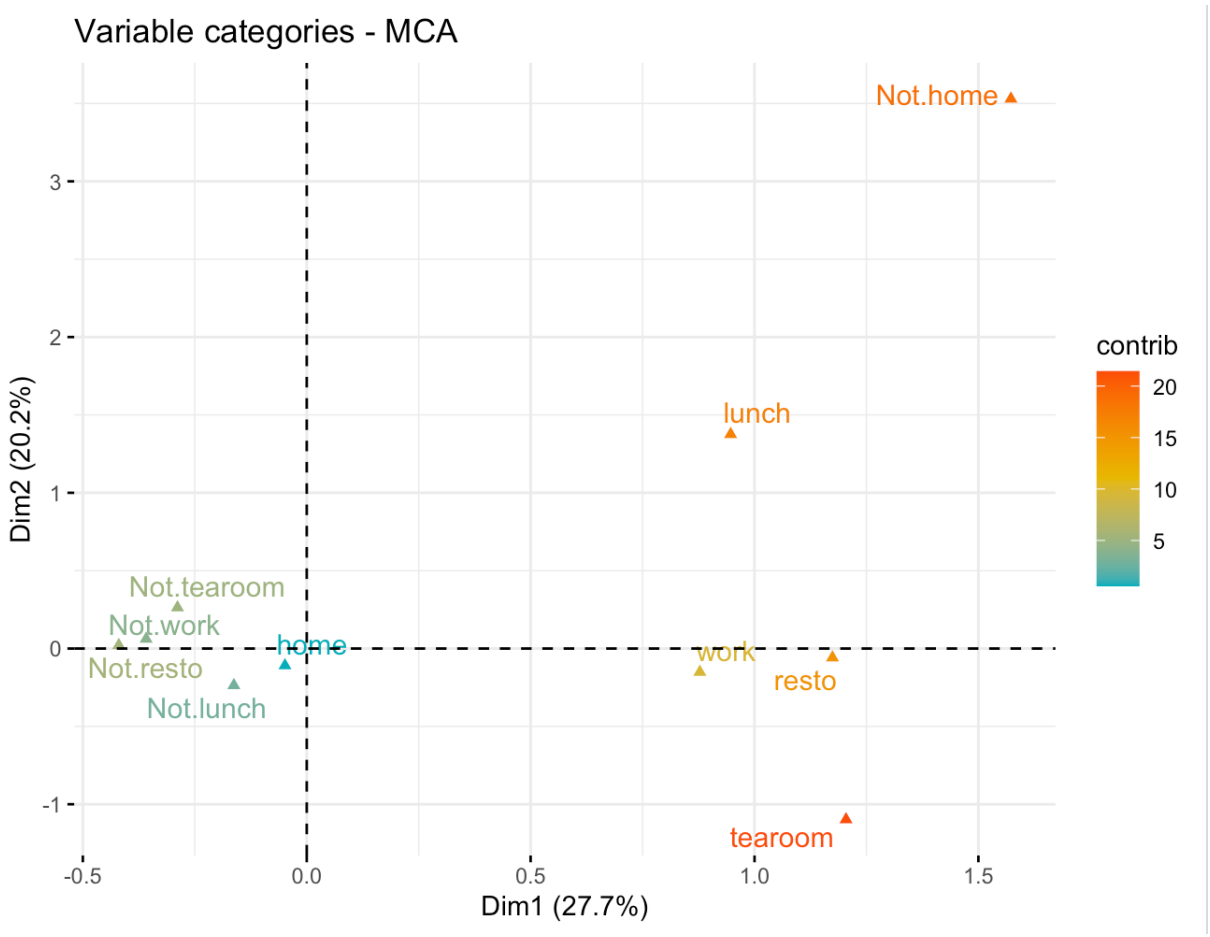


We can make the same conclusion, not.home, lunch and tearoom are the one that influence the most Dim2.

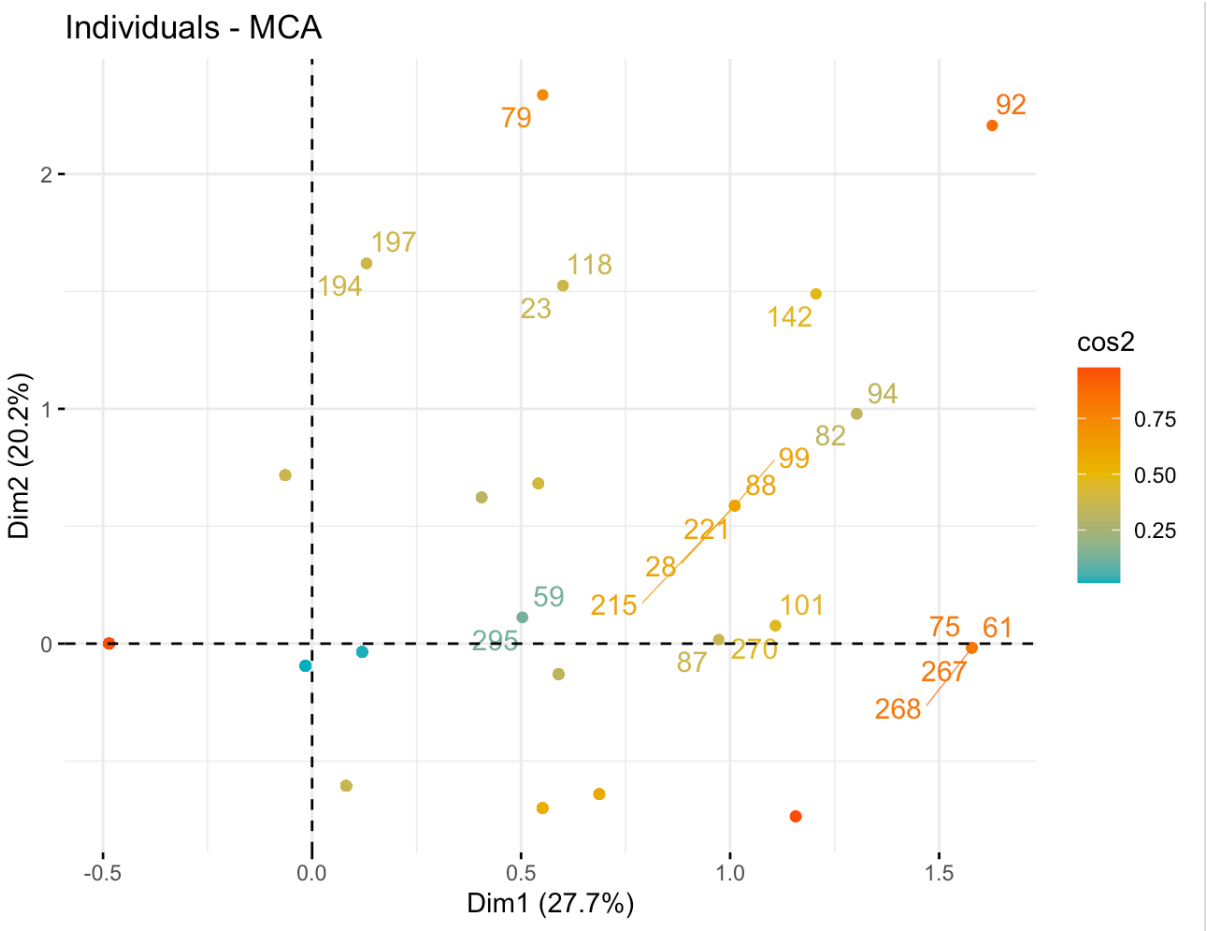
Finally, we can plot the contribution of variables on both Dim1 and Dim2 :



Which confirms that tearoom is the most influential variable, followed by not.home, lunch… Following the same process than for cos2, we have made a plot showing variables and their contribution percentage to both Dim1 and Dim2, which gives us a better visualization than the previous graphs.

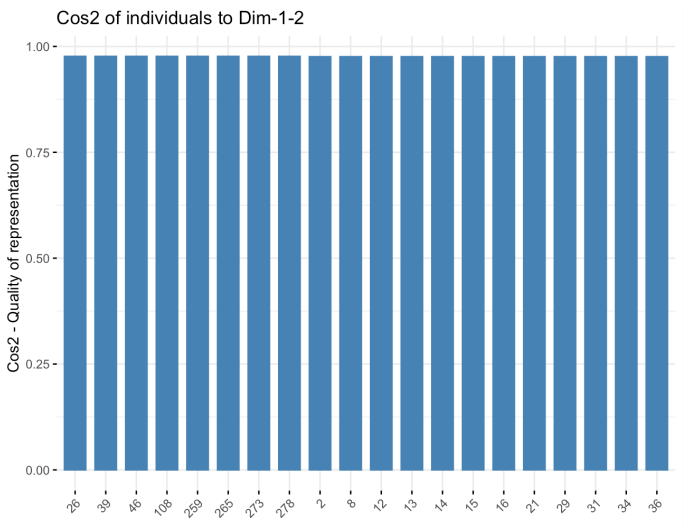


Let’s process the same way regarding individuals :

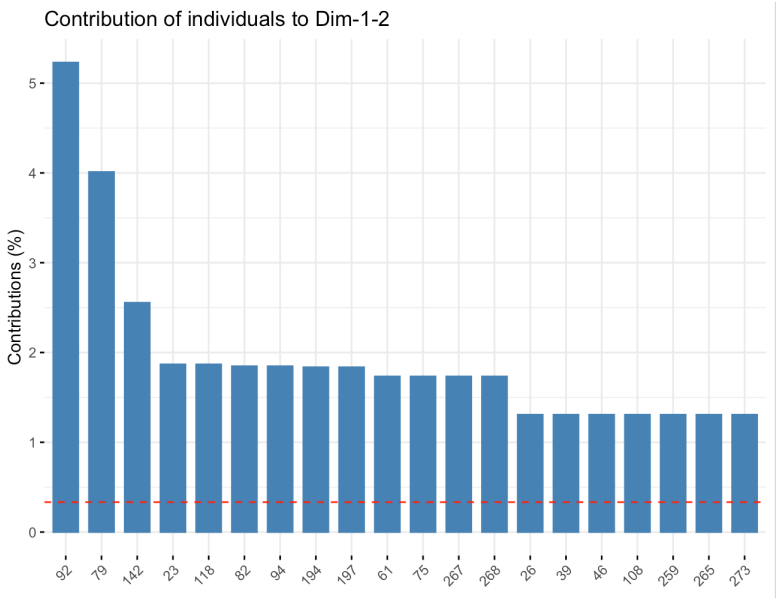


This plot is showing the cos2 distribution over the individuals of the dataset. This is a quite similar plot than we had earlier, where close points are quite similar. However it shows the influence of each rows on the dimensions by giving cos2, an indicator of their representation on the factor map. Points like on the bottom right which is not indexed are influencing a lot over Dim1 while 92, on the top right is influencing both Dim1 and Dim2 on a medium scale.

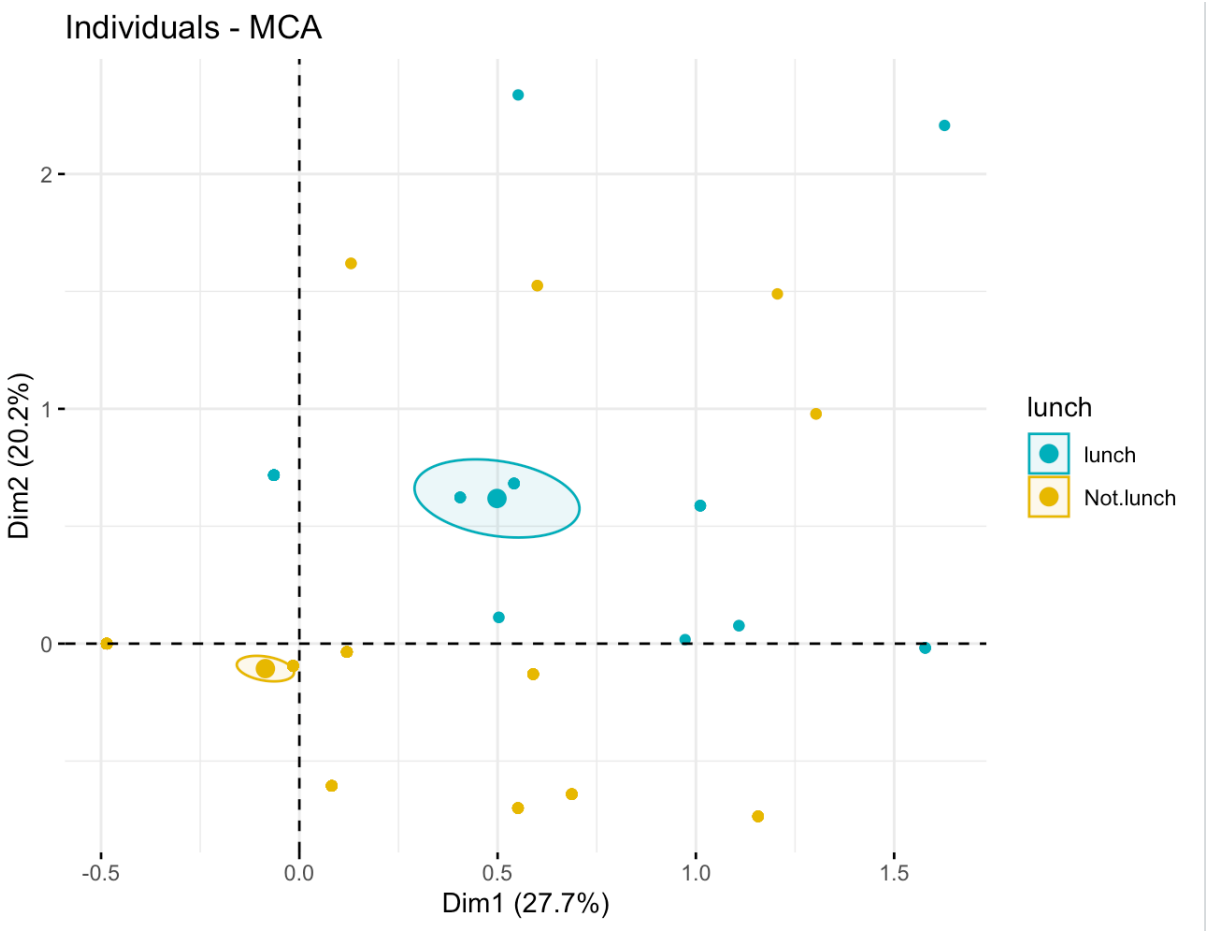
Here's the distribution of rows with the highest cos2 values over Dim1 and Dim2. Because we have a lot of rows (300), it is not really representative and doesn't display mid-influence values.



However, we can plot the distribution of the percentage of contribution of rows over Dim1 and Dim2, which can gives us more indications on the dataset :



We can see that those 20 rows are contributing to approximately 30% on Dim1 and Dim2. In other terms, 6,66% of the data are contributing for the third of our dimensions. The red line corresponds to the expected mean of the distribution if all the values were uniform. Close to 0.3, we can imagine that it is not really well distributed because some rows would contribute for almost nothing to the dimensions.



Finally, throughout our study we have plotted the distribution of the influence of the categorical variables over Dim1 and Dim2. The closer they are to each other, the more they are related.

With the different plot, the ellipses and our study in general we can answer our initial question and conclude that : lunch tea looks more associated with not drinking tea in a restaurant, at work or outside than in a tearoom.

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?

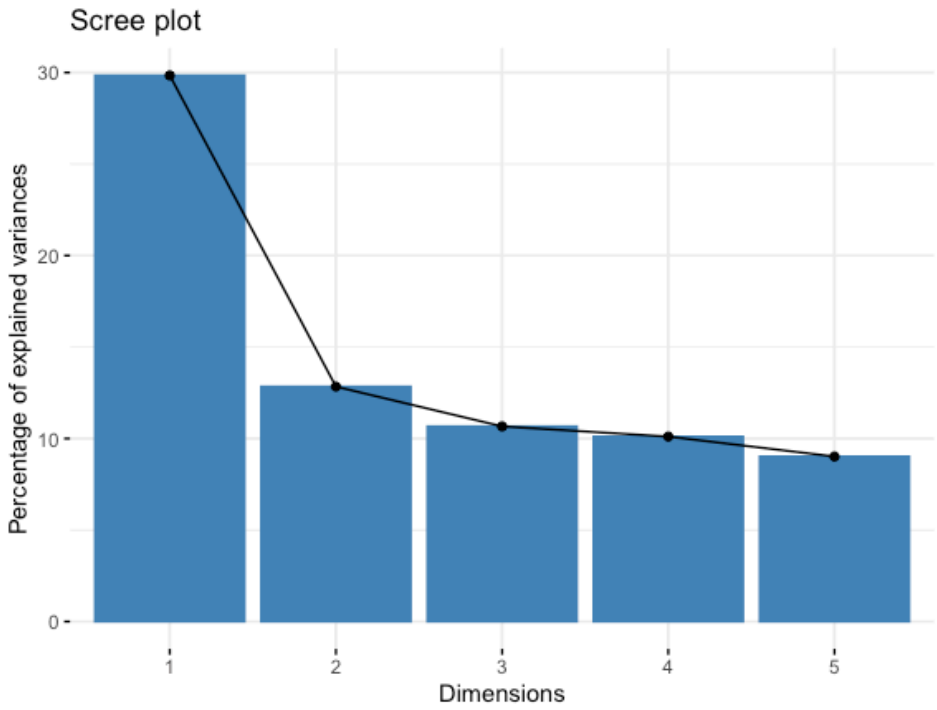
In this project, we would like to understand whether there is a link between characteristics of a student's life and how well they perform at school.

C – FAMD explained

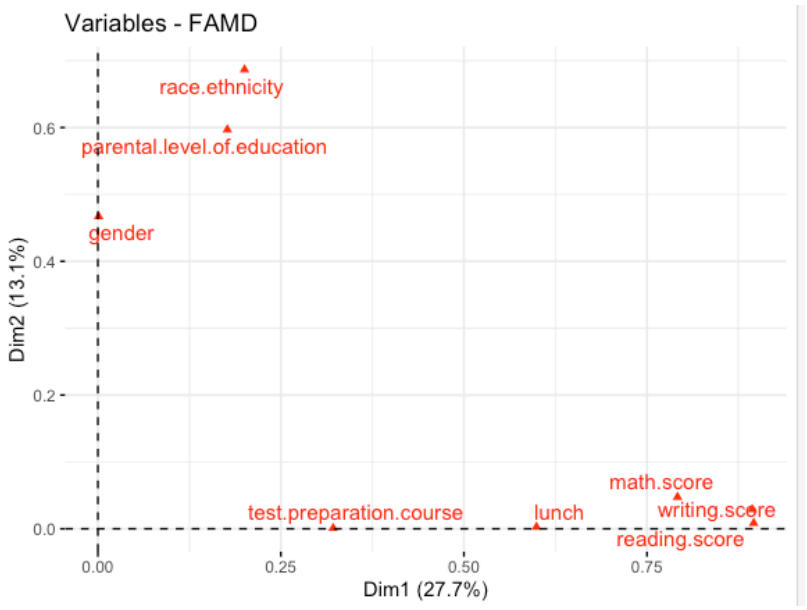
Factor Analysis of Mixed Data (FAMD) is a factorial method which is used on data tables which present both quantitative and qualitative variables. To treat the different type of data, FAMD works as PCA for quantitative variables and as MCA for qualitative variables. The methodology of MCA is explained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

D – Answering the question

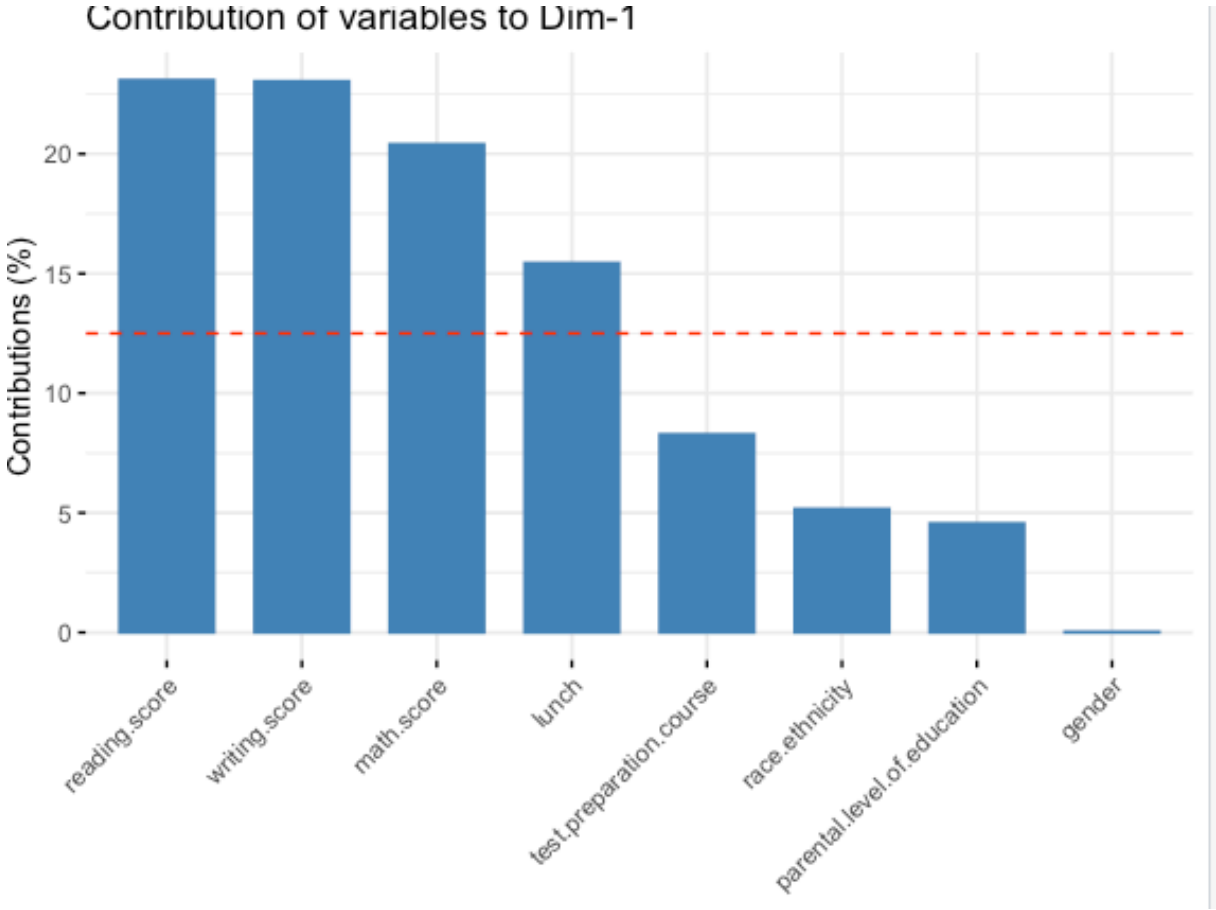
First of all, we upload the dataset in our code and use the 'as.factor' method to transform the categorical values. Here, it will be used for ethnicity, gender, lunch and preparation. We can then print the eigenvalues and display a graph to see how much each dimension describes the data.

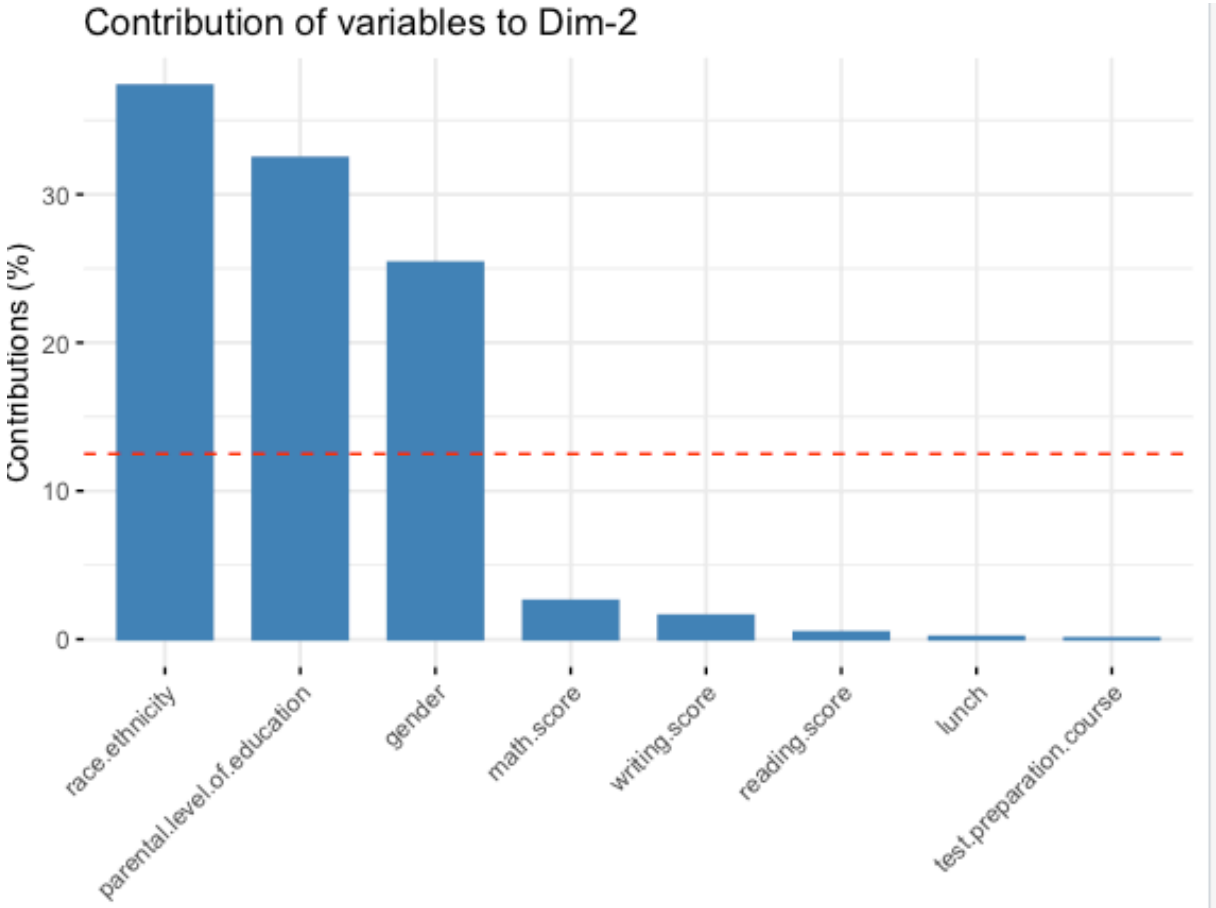


Here we get that the first two dimensions are describing 42.7% of the data set which isn't ideal but will be enough to observe correlations and links. Then we can plot the following graph that allows to see the contribution of each variable to the first two dimensions. We can see that parental level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

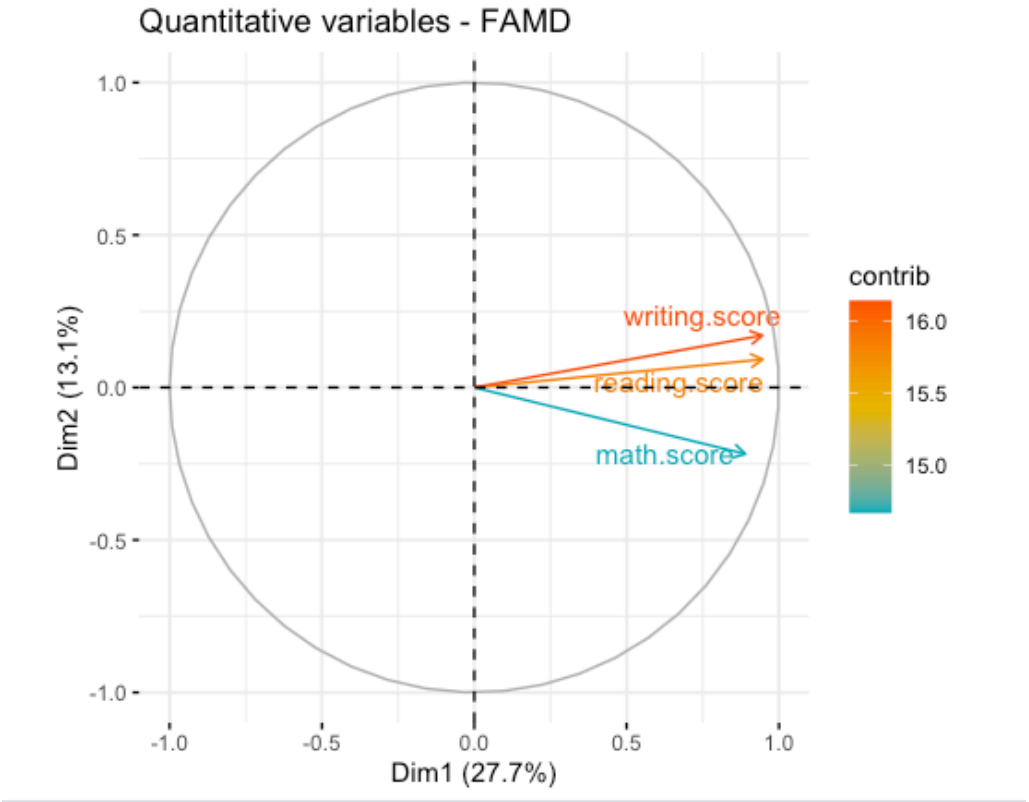


This can be observed more easily when looking at the following graphs of the contribution of each variable to each dimension. As we described above reading_score and writing_score contribute the most to dim1 and parental level of education and ethnicity to dimension 2.

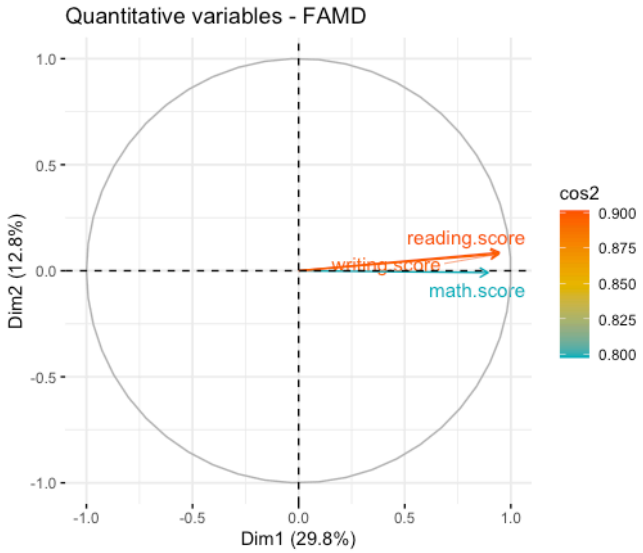

Contribution of variables to Dim-1
Contributions (%)
0
5
10
15
20
reading.score
writing.score
math.score
lunch
test.preparation.course
race.ethnicity
parental.level.of.education
gender


Contribution of variables to Dim-2
Contributions (%)
0
10
20
30
race.ethnicity
parental.level.of.education
gender
math.score
writing.score
reading.score
lunch
test.preparation.course

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.

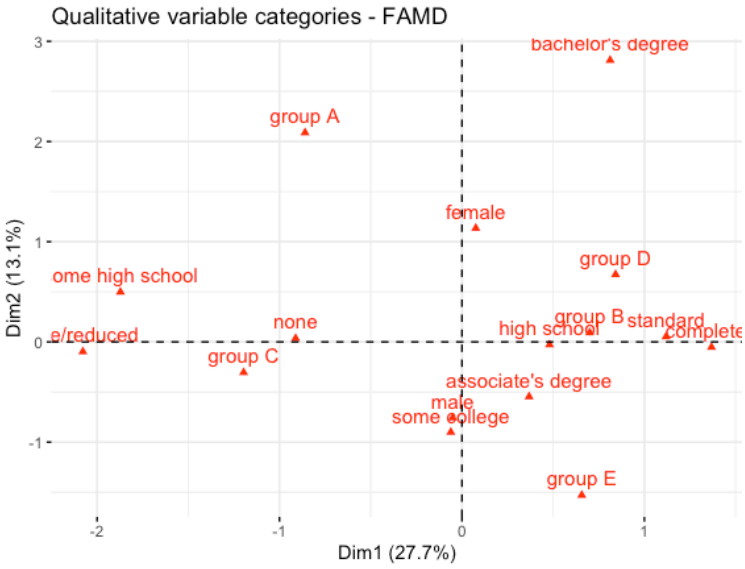


We can also get the cos2 score and plot here. Again, we can see that reading and writing score are well represented in the first two dimensions, but math score is not.

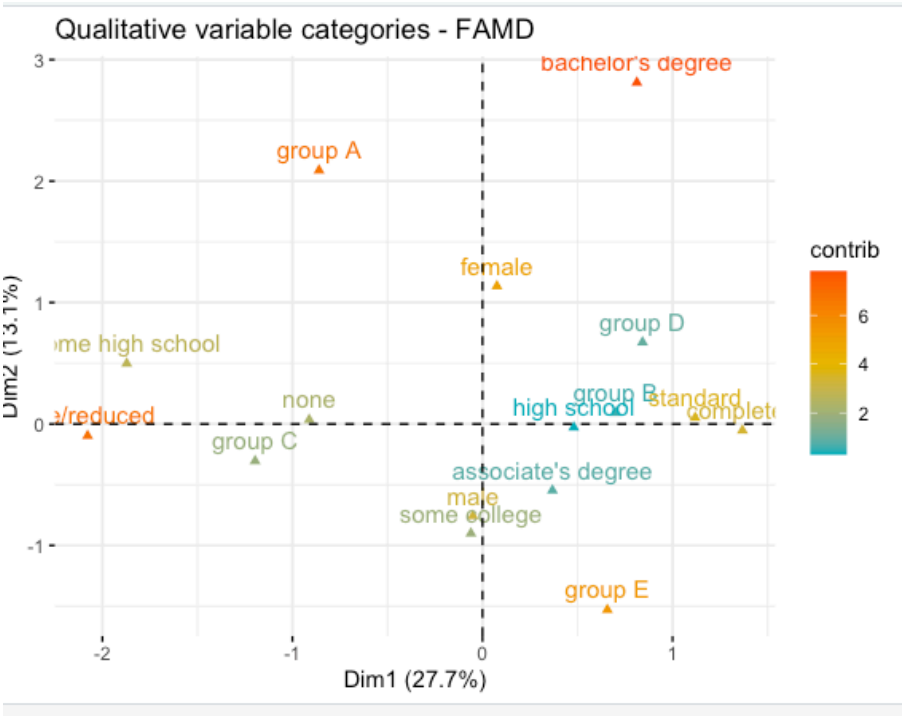


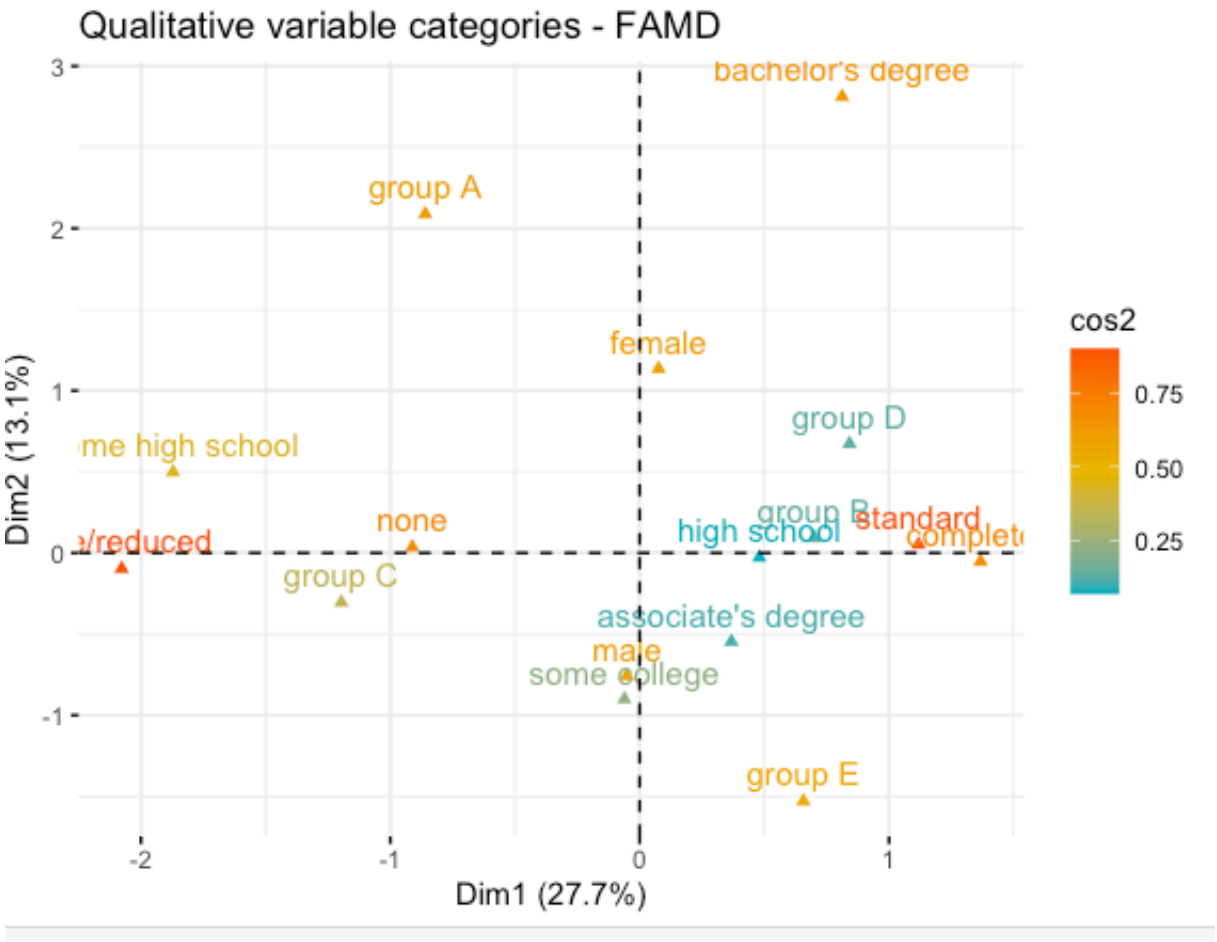
After that we can plot the qualitative variables and their different categories on the two-dimensional space. We can see that some variables are correlated and other are not. As an example, having parents

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.



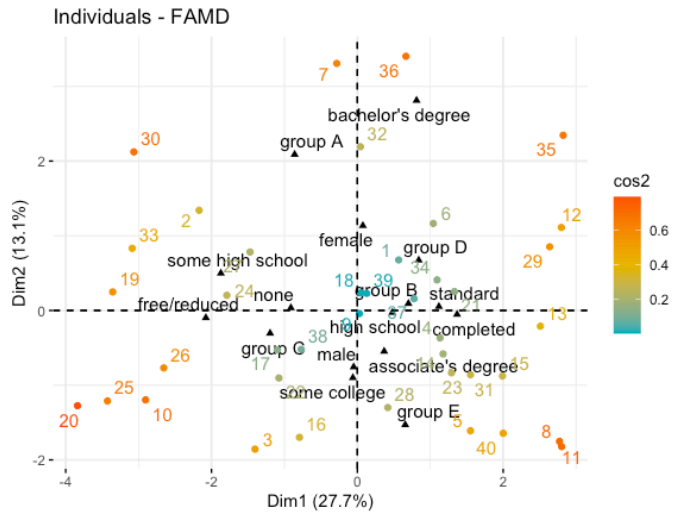
After that, we can see the contribution of the variables to the first dimensions, as well as their cos2 score. We can see that bachelor's degree, Group A and standard are well represented in the first 2 dimensions as an example, but high school and group D are not.

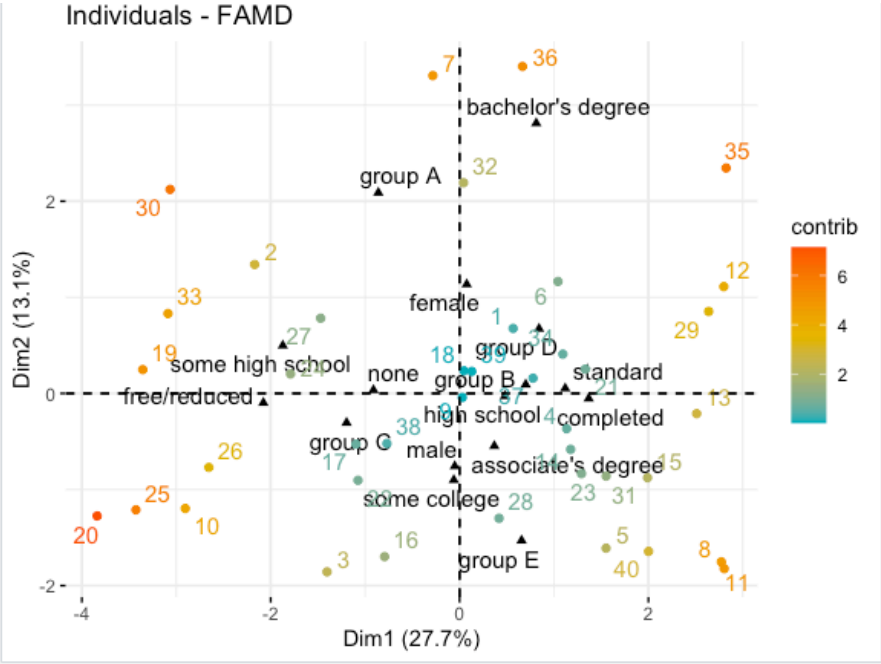




After looking at the different columns (variables), we can now look at the individuals and see their cos2 score as well as their contributions to the first 2 dimensions. We can see that some of the samples are much more accurately represented and contribute more to the two-dimensional space. These samples 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 are more correlated to higher grades with parents having a good level of studies than men are.