Homework 4

Practice of CA

1. Read the PCA_quetaltecaen data.
2. Perform a CA of this data. How many dimensions are significant?. Interpret the first factorial plan.
3. For the PCA_quetaltecaen data, compute the contribution of each cell to the total inertia, that is: $(fij - fi.\times f.j)^2/(fi.\times f.j)$. Compute the percentage of inertia due to the diagonal cells.
4. Clearly, the overloaded diagonal of the data set influences the results obtained (the overall inertia is mainly due to this overload diagonal). Try to nullify this influence by imputing the diagonal values by the independence hypothesis values of the product of marginal probabilities ($=n \times fi.\times f.j$). Take into account that each imputation modifies the marginal, hence you need an iterative algorithm.
5. Perform a new CA upon the quetaltecaen table, with the modified diagonal and interpret the results.

Practice MCA and Clustering

6. Read the file "mca_car.csv" containing the data and its dictionary about the cars and their characteristics found in specialized magazines. The final goal will be to find a model to predict the price of cars as function of its characteristics. First we will perform a visualization of the information contained in the dataset, then we will perform a clustering of cars. The data has been previously preprocessed to have it in categorical form.
7. With the obtained data frame perform a Multiple Correspondence Analysis. Take the brand and price (either categorical or continuous) as supplementary variables, whereas the remaining ones are active.
8. Interpret the first two obtained factors.
9. Decide the number of significant dimensions that you retain (by subtracting the average eigenvalue and represent the new obtained eigenvalues in a new screeplot).
10. Perform a hierarchical clustering with the significant factors, decide the number of final classes to obtain and perform a consolidation operation of the clustering.
11. Using the function *catdes* interpret and name the obtained clusters and represent them in the first factorial display.