

Homework 2

2024-11-18

Multiple Correspondence Analysis and Clustering

Import the data set “Wholesale customers data.csv” that composes of annual spending (in monetary units) of clients of a wholesale distributor on the products fresh, milk, grocery, frozen, detergents&paper and delicassen. The detailed information on the data set can be found in “<https://www.kaggle.com/datasets/binovi/wholesale-customers-data-set>”.

Our objective is to segment customers based on their spendings.

1. First do the exploratory data analysis.

- a) Import the data set correctly to R and assign type of each variable correctly. (0.5p)
- b) Convert the numerical variables to factors consist of two categories “low” and “high” by cutting them from their median and save them as new variables to the data set. (0.5p)

2. Application of Multiple Correspondence Analysis (MCA).

- c) Apply MCA on the categorical variables taking Region and Chanel as supplementary variables by using MCA() function in FactoMineR package. Interpret your findings. (3p)

3. Hierarchical Clustering based on MCA and PCA scores.

- d) Apply hierarchical clustering on MCA scores obtained from previous step by using HCPC function. How many clusters are constructed? (1.5p)
- e) Apply hierarchical clustering on PCA scores considering only the numerical variables and taking Region and Channel variables as supplementary variables. How many clusters are constructed? (1.5p)

4. Profiling

- f) Interpret the clusters obtained from both methods. (2p)
- g) Which of the above hierarchical clustering methods would you choose? Why? (1p)