RDSM 2022-2023

PROJET STATMULTIDIM / ACP/ ACM

Données de nutrition pour bébés et enfants

Un projet à faire à 2 ou 3, date limite de dépôt : 8 janvier

Il est recommandé de travailler ce projet en guise de préparation à l'examen.

Description du jeu de données : Le jeu de données est un extrait d'une base de données de composition alimentaire de plats préparés, biscuits, compotes, confiseries, boissons…
La base initiale contient 8618 lignes (plats) et 44 colonnes. On a filtré cette base pour ne retenir que les plats pour bébés (et pour enfants) avec les variables mesurant la composition en vitamines (en milligrammes, mg, ou microgrammes, mcg) et la composition nutritionnelle (graisse, protéines, sucre, fibre).

Le fichier RData « data.baby.RData » contient 2 tables, la première contenant les noms des plats ainsi que d'autres variables, l'autre table data.baby2.comp est celle qui sera analysée.

On se ramène donc à une table de 362 lignes et 13 colonnes.

Certaines variables de vitamines ont été renommées. Voici les correspondances :

Folate -> B9
Niacin -> B3
Riboflavin -> B2
Thiamin -> B1

L'objectif du projet est de rendre un document pdf ou html en RMarkdown. On cherche à comprendre ce qui caractérise la variabilité des plats et les différents types d'aliments, en se focalisant sur les vitamines.

Introduction

a) Pourquoi d'après-vous faut-il faire une ACP normée ?
b) Appliquez rapidement une ACP non normée : commentez la projection des variables et justifiez la position de la variable VitA_mcg.

A/ Analyse uni- et bivariées :

1° Analyse des variables

a) Calculez les valeurs moyennes et les quantiles 10% et 90% des variables numériques de data.baby2.comp
b) Donnez la matrice de corrélation et commentez.
c) Calculez avec ggpairs (package Ggally) la matrice des graphiques croisés pour les vitamines seulement. Justifiez le fait de supprimer dans la suite les valeurs extrêmes de VitB6, VitB12 et VitC (on recommande de supprimer de 1 à 10 plats).
d) Construisez la variable catégorielle CatSugar qui coupe la variables Sugar en 3 modalités selon les quartiles Q1 et Q3 puis tracer les boites à moustaches parallèles de VitC selon ces groupes d'aliments. Commenter.

B/ ACP normées du jeu de données

a) Faites l'ACP normée sur les variables de Vitamines en mettant les variables Fat, Sugar, Protein, Fiber en variables supplémentaires. Commentez l'éboulis des valeurs propres en analysant les pourcentages d'inerties, interprétez les axes, en commentant la projection des variables supplémentaires.

b) Donnez les 20 plats les plus contribuants à la construction de l'axe 1, à la construction de l'axe2, à la construction du plan (1,2).

c) Donnez une interprétation des positions de quelques plats sur le premier plan factoriel (et sur le second) **en vous aidant de leur intitul**é.

d) Commentez la position des plats 218 et 216, aidez-vous des quelques valeurs initiales données pour retrouver ces commentaires (selon les interprétations des axes).

e) Que vaut la variance des abscisses des points sur le premier plan factoriel ? donnez le calcul de 2 façons différentes.

C Inactivation des plats les plus extrêmes.

Reprenez l'analyse de la partie B (questions a) b), c)) en mettant certains plats en inactifs, pour cela :

a) Filtrez les données pour sélectionner tous les plats ayant une coordonnée :
   - Supérieure à 3 sur le premier axe
   - Négative sur le deuxième axe.

b) Reprenez l'ACP en mettant ces plats en **inactifs**. Commentez l'interprétation du nouveau plan factoriel (1,2).

D Analyse des correspondance multiples sur variables catégorisées.

a) Comme pour la variable de Sucre dans la partie A, pour toutes les variables de data.baby2.comp, contruisez des variables catégorielles à 3 modalités (dont les niveaux seront « Faible », « Moyen » et Fort »).
b) Applique l'ACM à ce tableau de données catégorielles en utilisant les seules mesures de vitamines comme variables actives (les autres étant inactives).
c) Retrouver l'inertie totale.
d) Tracer le graphe des modalités (sans projection des individus) et commentez les résultats.

ps : il est possible que la catégorisation en 3 modalités ne soit pas suffisantes, on aurait pu choisir de couper en 5.
pps : L'ACP est sensible aux données « extrêmes », l'ACM est sensibles aux modalités rares.