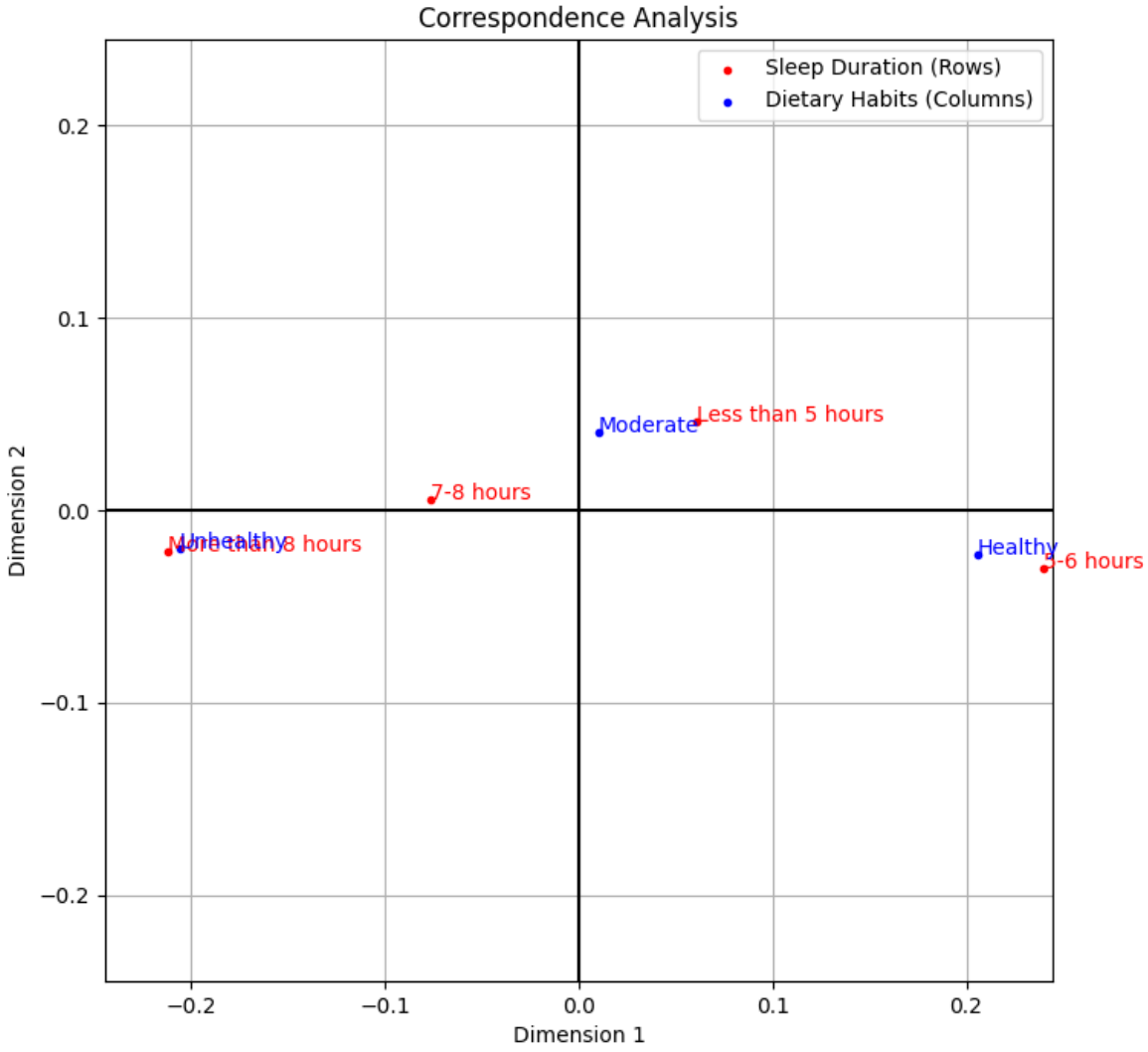
ENSEIRB-MATMECA

Département informatique

EI5IS102 - Traitement de l'information

RAPPORT DE PROJET

Etude d'un jeu de données à l'aide de l'AFC



Bourgouin Raphaël Dumarchat Joan Gouedard Anna

Année universitaire 2023-2024

Table des matières

Table des figures

1 Présentation de la méthode utilisée : AFC

L'Analyse Factorielle de Correspondance est une technique permettant d'analyser des données qualitatives. Plus précisément, elle permet d'analyser les relations entre 2 variables qualitatives catégorielles. Un représentation commune de ces données est la table de contingence, pour lequel un exemple est fourni dans la figure 1.

	Sleep duration				
Study satisafaction	Less than 5 hours	5-6 hours	7-8 hours	More than 8 hours	Total
1.0	23	19	20	24	86
2.0	25	25	25	25	100
3.0	19	25	33	26	103
4.0	29	29	31	27	116
5.0	27	25	19	26	97
Total	123	123	128	128	502

FIGURE 1 – Exemple de table de contingence

1.1 Notation

Dans la suite, on notera n le nombre total d'instances, V_1 la première variable (de taille I), V_2 la seconde (de taille J) et x_{ij} le nombre d'instances étant dans la catégorie i de la variable V_1 et dans la catégorie j de la variable V_2. On définit alors $X = (x_{ij})_{1\leq i\leq I, 1\leq j\leq J}$ la table de contingence. On peut alors parler des valeurs marginales des lignes et colonnes, dont les formules sont :

$$x_{i\bullet} = \sum_{j=1}^{J} x_{ij} \qquad x_{\bullet j} = \sum_{i=1}^{I} x_{ij}. \tag{1}$$

Cepandant, on préfèrera toujours travailler sur la table des probabilités définie par $f_{ij} = \frac{x_{ij}}{n}$, pour lesquelles on peut aussi définir les valeurs marginales avec :

$$f_{i\bullet} = \sum_{j=1}^{J} f_{ij} \qquad f_{\bullet j} = \sum_{i=1}^{I} f_{ij}. \tag{2}$$

1.2 Test d'indépendance

On souhaite dans un premier temps vérifier si les variables V_1 et V_2 sont indépendantes, ce qui est le cas si $\forall i, j, f_{ij} \approx f_{i,\bullet} f_{\bullet,j}$, on définit alors $\hat{f}_{ij} = f_{i,\bullet} f_{\bullet,j}$ la probabilité théorique sous l'hypothèse d'indépendance des variables. De manière similaire, on définit $\hat{x}_{ij} = n\hat{f}_{ij}$ les données théoriques sous l'hypothèse d'indépendance.

On peut alors procéder au test d'indépendance χ^2, qui consiste à :

- Calculer la distance $\chi^2_{obs} = \sum_{(i,j)} \frac{(x_{ij}-\hat{x}_{ij})^2}{\hat{x}_{ij}}$
- Fixer une p-value (usuellement à 0.05)
- Calculer le degré de liberté $df = (I-1)(J-1)$
- Déterminer $\chi^2_{critical}$ ou une p-valeur à l'aide d'une table
- Si $\chi^2_{obs} \leq \chi^2_{critical}$ ou p-valeur $\geq 5\%$ alors les variables sont indépendantes, sinon elles sont corrélées

1.3 Diagonalisation

Maintenant que l'on a confirmé que les variables sont corrélées, nous pouvons utiliser une technique plus précise afin d'obtenir plus d'informations, à savoir ici la diagonaliser la matrice des probabilités $F = (f_{ij})$.

On suppose ici qu'étudier F reviens à étudier $\tilde{F} = D_I F D_J$ avec $D_I = diag(\frac{1}{\sqrt{f_{1\bullet}}}, \ldots, \frac{1}{\sqrt{f_{I\bullet}}})$ et $D_J = diag(\frac{1}{\sqrt{f_{\bullet 1}}}, \ldots, \frac{1}{\sqrt{f_{\bullet J}}})$. On réalisera 2 diagonalisations, une sur $\tilde{F}\tilde{F}^T$ pour étudier les lignes et une sur $\tilde{F}^T\tilde{F}$ pour étudier les colonnes. On représente ensuite chaque catégorie de V_1 et V_2 dans un cercle des corrélations à la manière de l'ACP. Une propriété importante ici est le fait que les valeurs propres issues des 2 diagonalisations sont identiques, ce qui va nous permettre de représenter sur le même cercle les 2 variables. Dans ce graphique, si une catégorie de V_1 et de V_2 sont proches tout en étant éloignées de l'origine, cela montrera une corrélation entre cess deux catégories.

On pourra aussi calculer la contribution de chaque ligne/colonne dans les composantes principales afin de mieux interpréter les résultats.

2 Analyse des résultats

2.1 Etude du lien entre âge et habitudes alimentaires

2.1.1 Préparation des données

Nous avons ensuite voulu etudier le lien entre l'âge et les habitudes alimentaires. Cependant, la valeur du champ âge est directement l'âge, ce qui représente trop de catégories (une vingtaine) par rapport au nombre d'individus présent dans le jeu de données (500). Nous avons donc décidé de répartir les individus en tranches d'âges : les $18-22$, $22-26$, $26-30$ et $30+$. Nous pouvons maintenant voir dans le tableau de contingence en figure 2 que le nombre d'individus est suffisament élevé dans chaque catégorie pour pouvoir faire une AFC ayant du sens.

	Dietary Habits		
Age	Healthy	Moderate	Unhealthy
18-22	35	39	42
22-26	36	34	43
26-30	41	34	48
30+	49	65	36

FIGURE 2 – Table de contingence entre les habitudes alimentaires et les tranches d'âges

2.1.2 Test Chi-deux

Après exécution du test Chi-deux, la p-valeur obtenue est de $\approx 6\%$ ce qui et au-dessus de la p-valeur usuellement utilisée pour ce test, mais n'en est pas non plus très éloigné. Ainsi, nous avons quand même décidé de poursuivre l'analyse car cette p-valeur semble indiquer au moins une faible corrélation entre les deux variables.

2.1.3 Valeurs propres et cercle des corrélations

Après exécution de l'AFC, les deux composantes principales obtenues expliquent 100% de la variance, avec la première en expliquant $\approx 97\%$. Ainsi, la quasi totalité des corrélations seront montrées par la première composante, soit l'axe des abscisses du cercle des corrélations donné en figure 3.

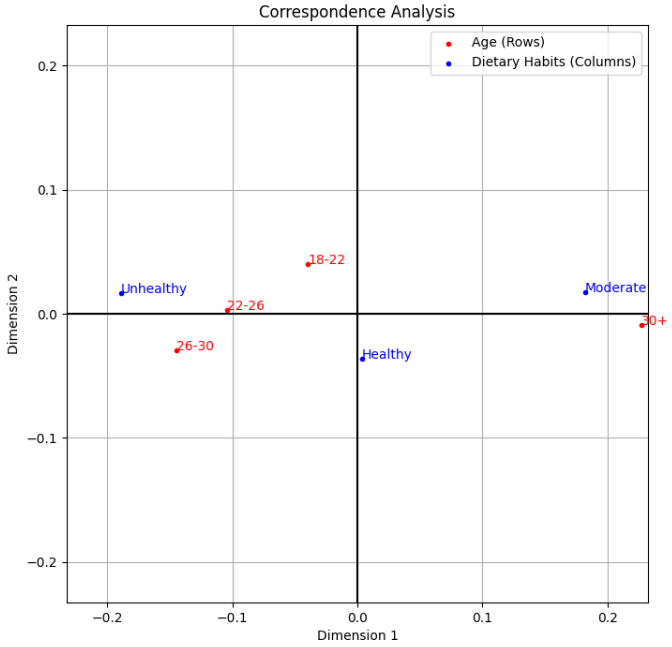


FIGURE 3 – Cercle des corrélations de l'AFC sur les tranches d'âges et les habitudes alimentaires

Sur la figure 3, on peut constater 2 légères tendances [1] :
— Les plus de 30 ans ont tendance à avoir une alimentation modérée
— Les 26-30 ans et 22-26 ans tendent quand à eux vers une alimentation plutôt mauvaise pour la santé

Cette tendance se confirme en regardant la contribution de chaque catégorie sur les composantes principales. En effet, comme on peut le voir sur la figure 4, la première composante dépend des catégories Moderate et Unhealthy de manière approximativement égale, et la figure 3 montre en complément que Moderate y contribue positivement et Unhealthy y contribue négativement. Ainsi, l'analyse de correspondance a classé les tranches d'ages selon le nombre d'individus ayant un régime modéré moins le nombre d'individus ayant un régime mauvais pour la santé.

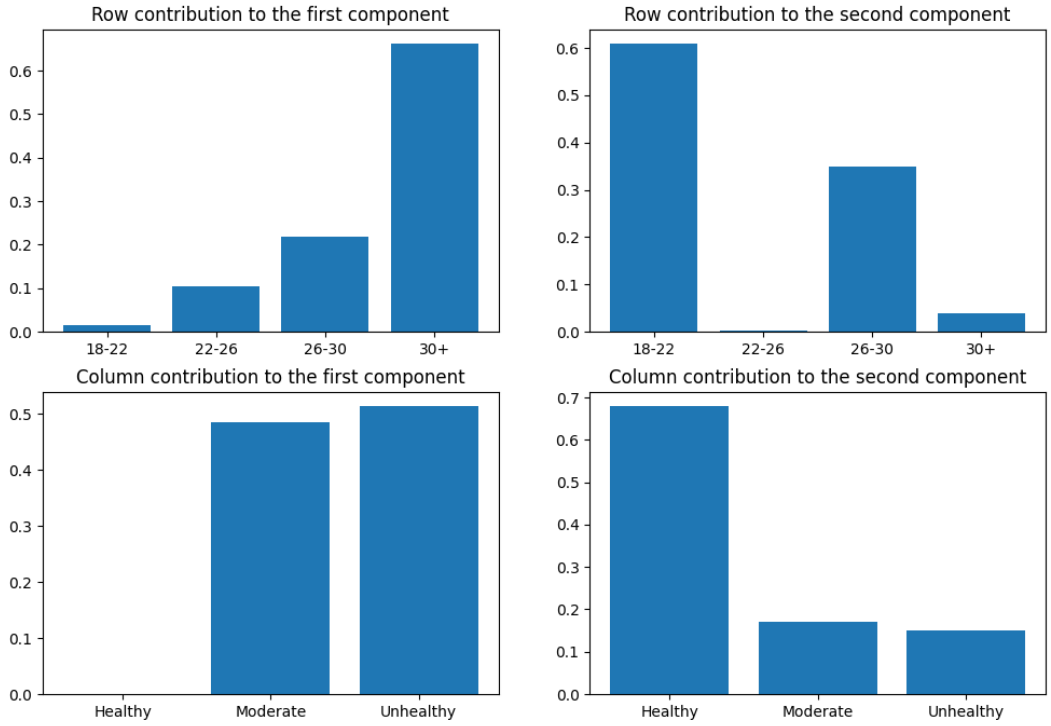


FIGURE 4 – Contribution des catégories tranches d'âges et habitudes alimentaires dans les composantes principales

3 Test

[4, 2, 3, 1]

Références

[1] Bibliotèque python matplotlib. *https://matplotlib.org/*.

[2] Bibliotèque python prince. *https://pypi.org/project/prince/*.

[3] Bibliotèque python scipy. *https://scipy.org/*.

[4] Depression student dataset.
https://www.kaggle.com/datasets/ikynahidwin/depression-student-dataset.

1. Nous insistons vraiment sur le fait que ces corrélations sont faibles et ne représentent qu'au plus des légères tendances