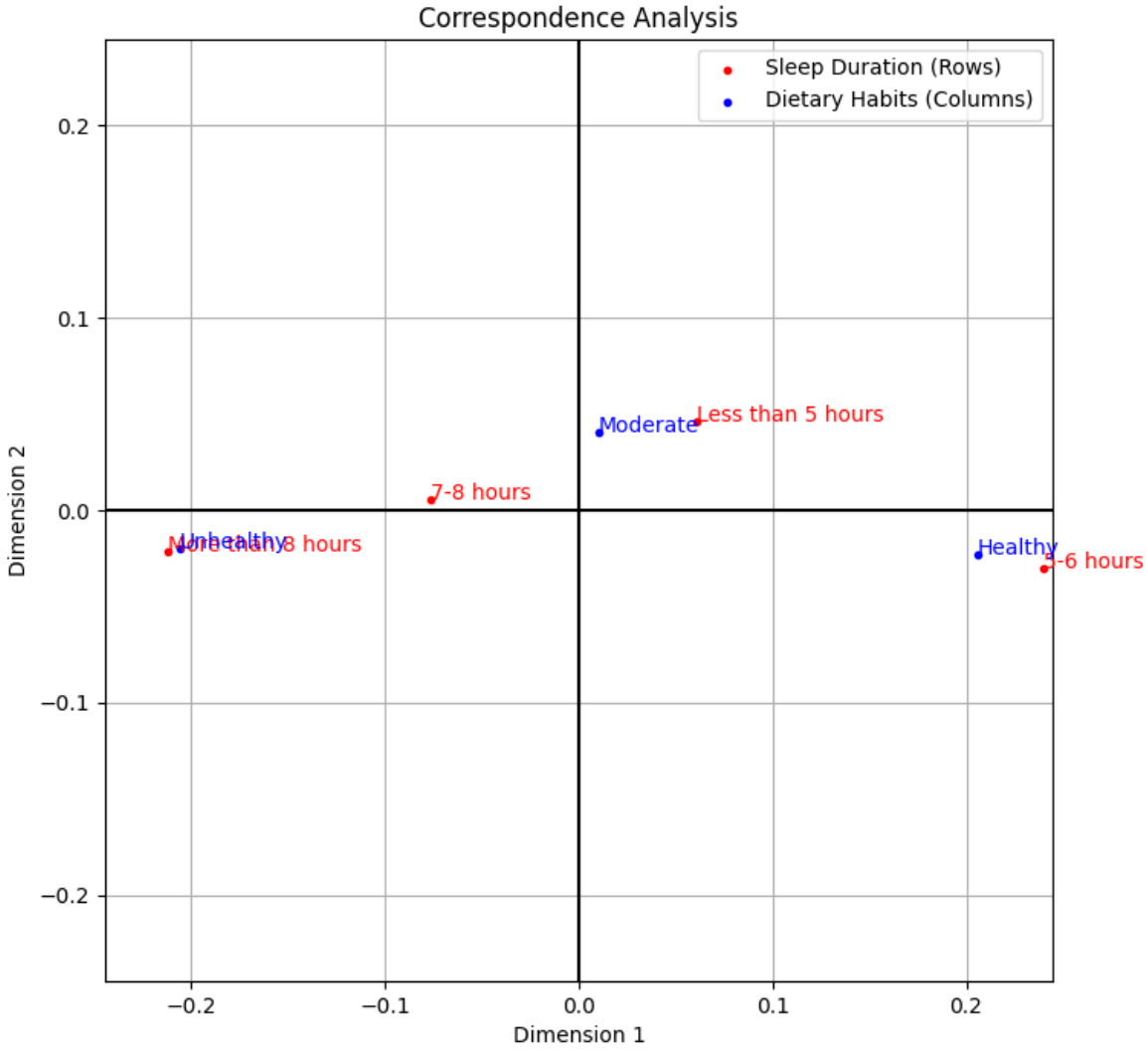
ENSEIRB-MATMECA

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 Introduction

Le présent rapport étudie des données publiques portant sur la dépression chez les étudiants [9]. Afin de traiter et analyser ces données, nous avons choisis l'Analyse Factorielle des Correspondances (abrégé en AFC), les données étant categorielles. Notre objectif est d'identifier des relations entre différentes catégories, selon si l'on regarde les individus dépréssifs, les individus non dépréssifs ou tous les individus. Ainsi, on pourra potentiellement constater certains effets de la dépréssion à travers les différences de corrélations entre dépréssifs et non dépréssifs.

Nous commencerons par présenter le jeu de données que nous utilisons en section 2. Ensuite nous présenterons la méthode utilisée dans ce rapport, l'AFC, en section 3. Finalement, nous analyserons les résultats obtenus dans la section 4 avant de conclure en section 5.

2 Jeu de données et outils utilisés

Nous avons choisi d'étudier le jeu de données "Depression Student Dataset" [9], constitué de données académiques (pression académique par exemple), économiques (stress financier par exemple), ainsi que sur des habitudes de vie (temps de sommeil par exemple) recueillie sur 502 individus, la moitée étant dépréssifs.

Nous avons pour l'analyse utilisé le language Python [1] sur un jupyter notebook [2]. De plus, nous avons utilisé les librairies pandas [3] afin d'importer et gérer les données, matplotlib [4] pour tracer les graphiques, prince [5] afin de réaliser l'AFC ainsi que scipy [6], qui nous a permis de réaliser les tests chi-deux.

Le code écrit dans le cadre de notre projet est trouvable sur le dépôt github du projet [7].

3 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

3.1 Notations

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 :

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

3.2 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

3.3 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 $D_I = diag(\frac{1}{\sqrt{f_{1\bullet}}}, \dots, \frac{1}{\sqrt{f_{I\bullet}}})$ et $D_J = diag(\frac{1}{\sqrt{f_{\bullet 1}}}, \dots, \frac{1}{\sqrt{f_{\bullet J}}})$. 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.

4 Analyse des résultats

4.1 Etude du lien entre âge et habitudes alimentaires

4.1.1 Préparation des données

Nous avons ensuite voulu etudier le lien entre l'âge et les habitudes alimentaires. Cepandant, 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

4.1.2 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.

4.1.3 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.

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

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

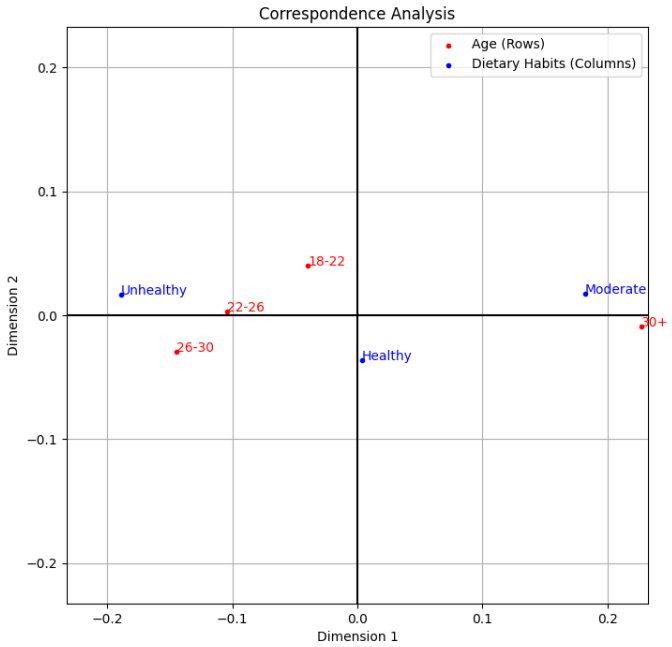


FIGURE 3 – Cercle des corrélations de l'AFC sur les tranches d'âges et les habitudes alimentaires

4.1.4 Etude sur les données seulement avec dépréssifs et sans dépréssifs

Lorsque que nous réalisons le test chi-deux sur le jeu de données sans les individus dépréssifs et avec seulement les individus dépréssifs, la p-valeur obtenue est respectivement de $\approx 51\%$ et de $\approx 42\%$, ce qui indique que les variables tranches d'âge et habitudes alimentaires sont indépendantes. Ceci peut être surprenant, étant donné qu'en faisant l'analyse avec la totalité des données on obtient une faible corrélation. Il semblerait ici que la cause soit un effet similaire au paradoxe de Simpson [8], où une troisième variable est affectée par notre choix de séparation de la population, causant cette décorrélation. Cependant cela reste à confirmer.

5 Conclusion

Références

[1] *https://www.python.org*.

[2] *https://jupyter.org/*.

[3] *https://pandas.pydata.org/*.

[4] *https://matplotlib.org/*.

[5] *https://pypi.org/project/prince/*.

[6] *https://scipy.org/*.

[7] *https://github.com/bourgouinraphael/Projet_traitement_info/blob/main/Projet.ipynb*.

[8] *https://fr.wikipedia.org/wiki/Paradoxe_de_Simpson*.

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