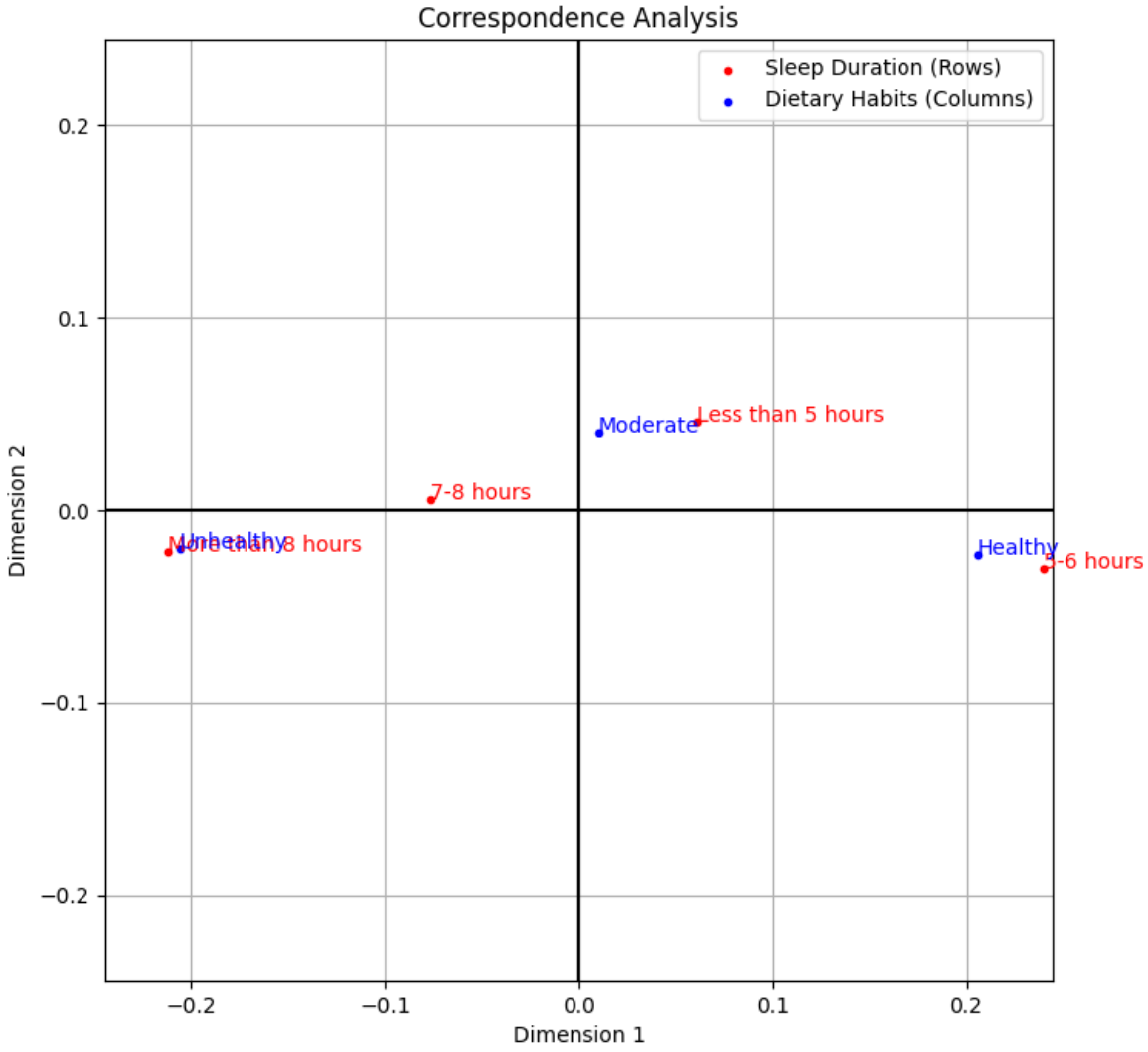
ENSEIRB-MATMECA

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

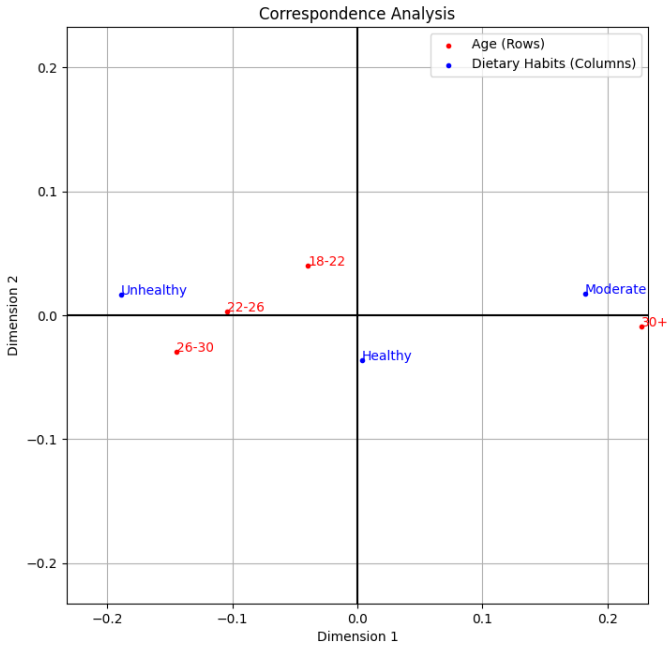


FIGURE 3 – Cercle des corrélations de l'AFC sur les tranches d'âges et les habitudes alimentaires

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.

4.2 Etude du lien entre le temps d'études et le sommeil

4.2.1 Tableau de contingence et test chi-deux pour l'ensemble des individus

Nous avons aussi voulu étudier le lien entre le temps que consacre les étudiants pour leurs études par jour et leurs temps de sommeil. Pour commencer nous pouvons voir avec le tableaux de contingence en figure 4 que si on prends l'ensemble des individus alors la répartition dans chaque catégorie est suffisante pour avoir une analyse cohérente.

	sleep duration			
study hours	5-6 hours	7-8 hours	Less than 5 hours	More than 8 hours
3-	20	28	25	27
3-6	22	24	31	29
6-9	24	31	26	42
9+	57	45	41	30

FIGURE 4 – Table de contingence entre le temps consacré par jour aux études et le temps de sommeil

Nous avons ensuite réalisé un test Chi-deux pour être sûre que l'AFC est du sens. La p-valeur obtenue est de $\approx 3,2\%$ ce qui est suffisant pour montré une corrélation entre les deux valeurs.

4.2.2 Tableau de contingence et test chi-deux pour les individus dépressifs et non dépressifs

Si on limite l'analyse seulement aux individus non dépressifs on obtient la table de contingence en figure 5. Cette limitation fait disparaitre la corrélation entre les deux variables. En effet, la p-valeurs est de $\approx 13,6\%$ ce qui est bien supérieur aux valeurs nécessaires pour qu'une corrélation soit observé.

	sleep duration			
study hours	5-6 hours	7-8 hours	Less than 5 hours	More than 8 hours
3-	12	18	19	18
3-6	16	10	16	16
6-9	7	14	13	20
9+	24	19	11	17

FIGURE 5 – Table de contingence entre le temps consacré par jour aux études et le temps de sommeil par les individus non depressifs

Ensuite, si on limite seulement aux individus dépressifs on obtient le tableau de contingence en figure 6. On voit que de la même que pour les individus non dépressifs la corrélation disparait. La p-valeurs est de $\approx 5,8\%$ ce qui est assez proche des valeurs prisent pour indiquer une faible corrélation qui n'est pas aussi forte qu'avec tous les individus sans distinctions.

	sleep duration			
study hours	5-6 hours	7-8 hours	Less than 5 hours	More than 8 hours
3-	8	10	6	9
3-6	6	14	15	13
6-9	17	17	13	22
9+	33	26	30	13

FIGURE 6 – Table de contingence entre le temps consacré par jour aux études et le temps de sommeil par les individus depressifs

Pour résumer, il n'y pas de corrélation entre le temps consacré aux études et la durée du sommeil lorsque nous observons un groupe restreint d'individu (dépressifs et non dépressifs). Nous observons donc un inversement de la tendance lorsque nous prennons les individus dans leur totalité. Ce phénomène s'apparente encore une fois au paradoxe de Simpson [8]. Il peut s'expliquer par une troisième variable affectée par ce choix de séparé les individus dépressifs et non dépressifs.

4.2.3 Valeurs propres et cercle de corrélations

Après avoir réalisé l'AFC, on obtient le cercle de corrélation en figure 7. Nous trouvons que les deux composantes principales explique $\approx 96,5\%$ de la variance avec en particulier la première qui en explique $\approx 84,9\%$. Il faut donc regardé en priorité l'axe des abscisses pour chercher des corrélations.

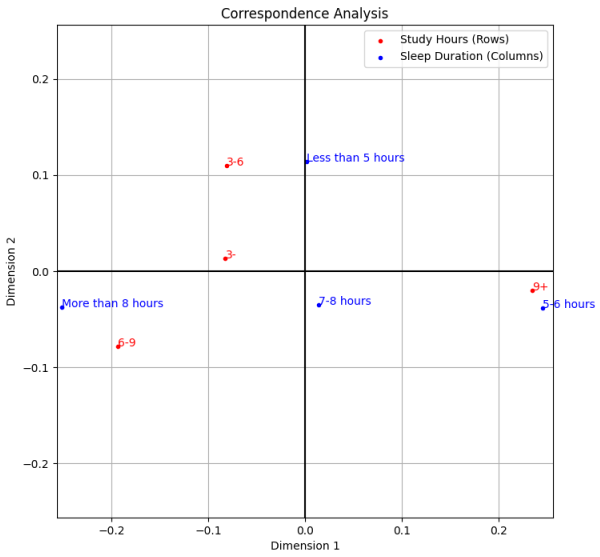


FIGURE 7 – Cercle des corrélations de l'AFC sur les tranches d'âges et les habitudes alimentaires

C'est ainsi que nous remarquons deux tendances. En effet, nous remarquons que les étudiants travaillant le plus, c'est à dire 9 heures ou plus, ont tendance à dormir entre 5 et 6 heures. Cette tendance s'explique facilement, les étudiants travaillant beaucoup passe moins de temps à dormir car ils utilisent leurs temps pour travailler. Ces étudiants n'ont pas les durées de sommeil les plus courtes non plus car ce comportement corresponds à des étudiants passant beaucoup de temps à travailler mais aussi à des étudiants ayant des comportement différentes ou étant dans des situations particulières comme par exemples des étudiants victime d'insomnie.

Le cercle de corrélation montre aussi une deuxième tendance, celle-ci un plus faible que la première. En effet, on remarque que les étudiants dormant plus de 8h ont tendance à travailler entre 6 et 9 heures. Ces étudiants travaille beaucoup, deuxième catégorie avec la plus grosse durée de travail, mais ont aussi le temps de sommeil le plus long. Cela peut s'expliquer par des étudiants gérant mieux leurs temps en travaillant beaucoup sans avoir à perdre du temps de sommeil.

Cette tendance est confirmé par la figure 8. Nous remarquons qu'en accord avec la figure 7 les catégories contribuant le plus à la première composante sont 5-6 hours et More than 8 hours de manière approximativement égale pour le temps de sommeil. Nous retrouvons pour le temps consacré aux études que les catégories contribuant le plus sont bien 9+ et 6-9, mais avec un écart bien plus important entre les contribution de ces catégories. Ensuite, pour la deuxième composante nous remarquons la catégorie 6-9 qui a une contribution assez élévé et nous retrouvons aussi des catégories dont nous n'avons pas tiré de corrélation. Ce qui s'explique par le fait que la seconde composante n'explique que $\approx 11,6\%$ de la variance.

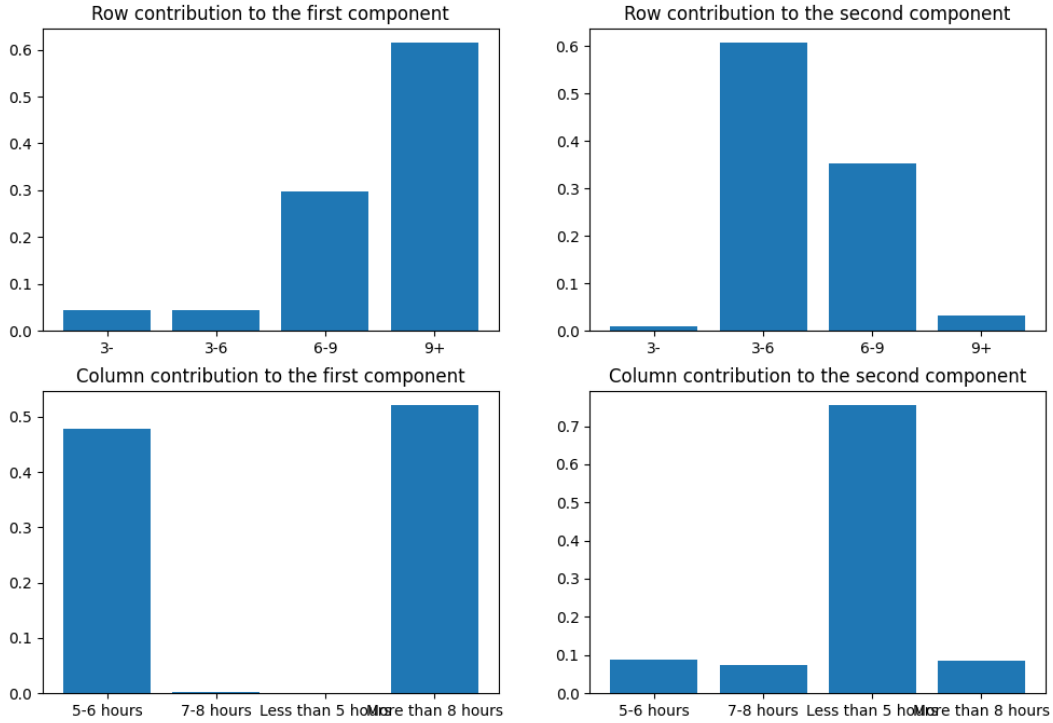


Figure 8 – Contribution des catégories dans les composantes principales

5 Conclusion

Pour conclure, L'AFC de cette base de donnée nous à permit de mettre en valeur différente corrélation. Notamment entre la tranche d'age et le régime alimentaire, le temps consacré aux études et le temps de sommeil ou le temps de sommeil et le régime alimentaire. nous avons aussi montré que les résultats pouvait différé lorsque nous séparons les individus dépressifs et non dépressifs. Nous avons notamment mis en évidence des phénomènes s'apparentant au paradoxe de Simpson.

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.