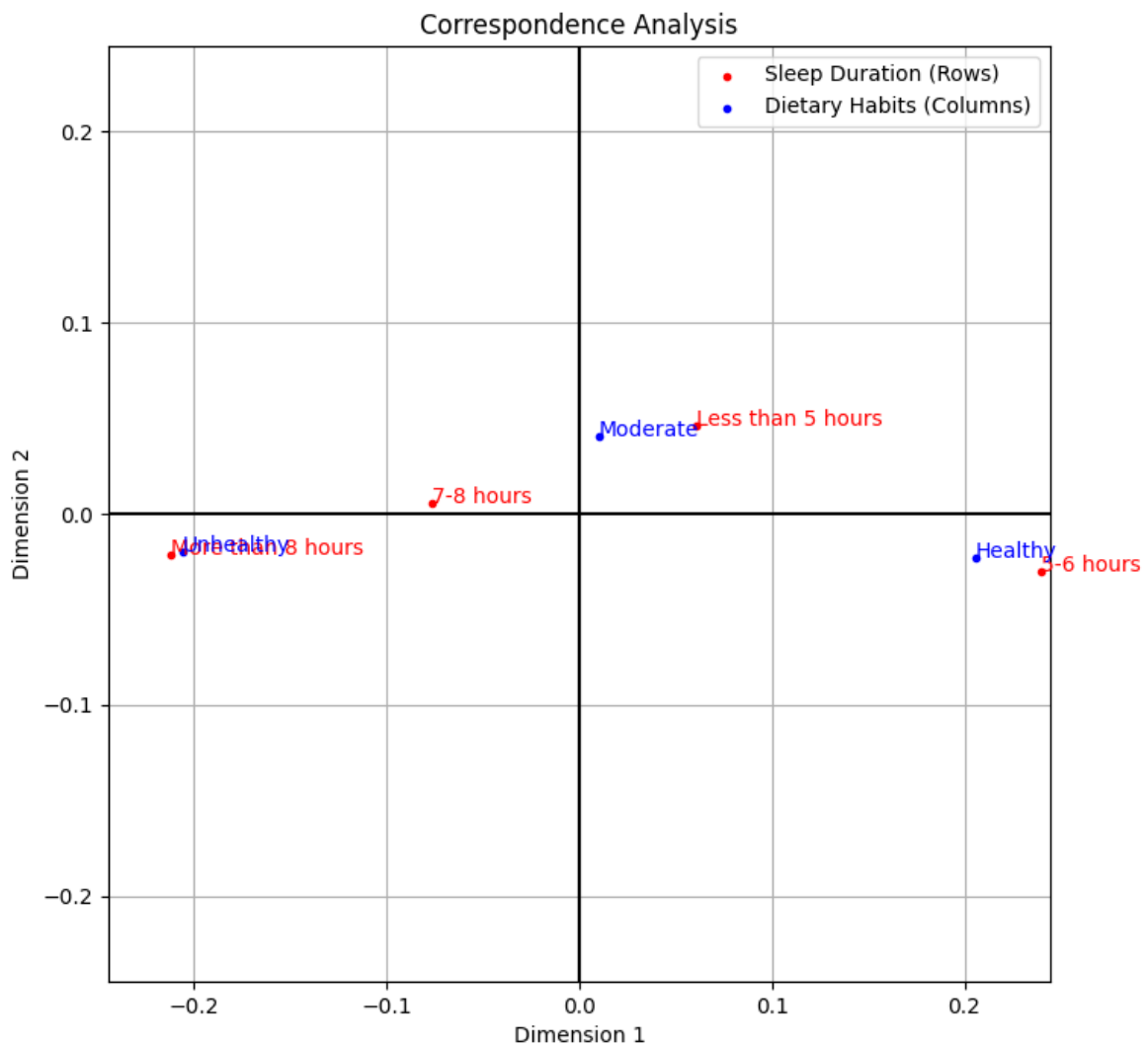


RAPPORT DE PROJET

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

1	Introduction	1
2	Jeu de données et outils utilisés	1
3	Présentation de la méthode utilisée : AFC	1
3.1	Notations	1
3.2	Test d'indépendance	1
3.3	Diagonalisation	2
4	Analyse des résultats	2
4.1	Etude du lien entre âge et habitudes alimentaires	2
4.1.1	Préparation des données	2
4.1.2	Test Chi-deux	2
4.1.3	Valeurs propres et cercle des corrélations	2
4.1.4	Etude sur les données seulement avec dépressifs et sans dépressifs	3
4.2	Etude du lien entre le temps d'études et le sommeil	3
4.2.1	Tableau de contingence et test chi-deux pour l'ensemble des individus	3
4.2.2	Tableau de contingence et test chi-deux pour les individus dépressifs et non dépressifs	3
4.2.3	Valeurs propres et cercle de corrélations	4
4.3	Etude du lien entre temps de sommeil et habitudes alimentaires	5
4.3.1	Tableau de contingence et test chi-deux pour l'ensemble des individus	5
4.3.2	Test chi-deux sur les individus dépressifs et non dépressifs	6
4.3.3	Etude des cercles de corrélation	6
4.3.4	Contributions des variable dans les composantes principales	7
5	Conclusion	8

Table des figures

1	Exemple de table de contingence	1
2	Table de contingence entre les habitudes alimentaires et les tranches d'âges	2
3	Cercle des corrélations de l'AFC sur les tranches d'âges et les habitudes alimentaires	3
4	Table de contingence entre le temps consacré par jour aux études et le temps de sommeil	3
5	Table de contingence entre le temps consacré par jour aux études et le temps de sommeil par les individus non depressifs	4
6	Table de contingence entre le temps consacré par jour aux études et le temps de sommeil par les individus depressifs	4
7	Cercle des corrélations de l'AFC sur le temps d'études et le temps de sommeil	4
8	Contribution des catégories de temps d'études et temps de sommeil dans les composantes principales	5
9	Table de contingence entre le temps de sommeil et les habitudes alimentaires	5
10	Cercle des corrélations de l'AFC entre le temps de sommeil et les habitudes alimentaires sur l'ensemble des individus	6
11	Cercles des corrélations de l'AFC entre le temps de sommeil et les habitudes alimentaires pour les dépressifs (à gauche) et les non dépressifs (à droite)	7
12	Contributions des catégories de temps de sommeil et de qualité d'alimentation dans composantes principales pour l'ensemble des individus	7
13	Contributions des catégories de temps de sommeil et de qualité d'alimentation dans composantes principales pour les dépressifs (à gauche) et les non dépressifs (à droite)	8

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 catégorielles. Notre objectif est d'identifier des relations entre différentes catégories, selon si l'on regarde les individus dépressifs, les individus non dépressifs ou tous les individus. Ainsi, on pourra potentiellement constater certains effets de la dépression à travers les différences de corrélations entre dépressifs et non dépressifs.

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 moitié étant dépressifs.

Nous avons pour l'analyse utilisé le langage Python [1] sur un jupyter notebook [2]. De plus, nous avons utilisé les bibliothèques 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. Une représentation commune de ces données est la table de contingence, pour lequel un exemple est fourni dans la figure 1.

	Sleep duration				
Study satisfaction	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 :

$$x_{i\bullet} = \sum_{j=1}^J x_{ij} \quad x_{\bullet j} = \sum_{i=1}^I x_{ij}. \quad (1)$$

Cependant, on préférera 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 :

$$f_{i\bullet} = \sum_{j=1}^J f_{ij} \quad f_{\bullet j} = \sum_{i=1}^I f_{ij}. \quad (2)$$

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 $\hat{x}_{ij} = n f_{ij}$ les données théoriques sous l'hypothèse d'indépendance.

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

- Calculer la distance $\chi_{obs}^2 = \sum_{(i,j)} \frac{(x_{ij} - \hat{x}_{ij})^2}{\hat{x}_{ij}}$
- Fixer une p -value (usuellement à 0.05)
- Calculer le degré de liberté $df = (I - 1)(J - 1)$
- Déterminer $\chi_{critical}^2$ ou une p -valeur à l'aide d'une table
- Si $\chi_{obs}^2 \leq \chi_{critical}^2$ 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 revient à étudier $\tilde{F} = D_I F D_J$ avec $D_I = \text{diag}(\frac{1}{\sqrt{f_{1\bullet}}, \dots, \frac{1}{\sqrt{f_{I\bullet}}})$ et $D_J = \text{diag}(\frac{1}{\sqrt{f_{\bullet 1}}, \dots, \frac{1}{\sqrt{f_{\bullet J}}})$.

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 ces 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 étudier 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 suffisamment 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 est 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 tendances¹ :

- 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

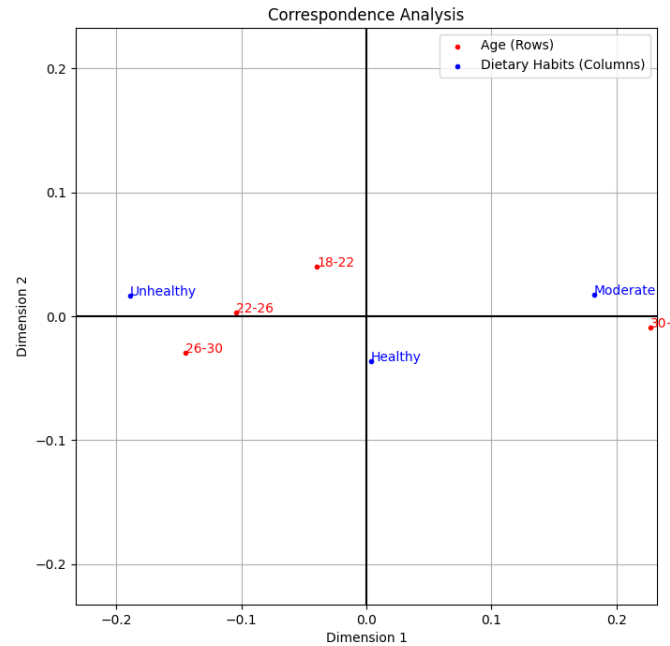


FIGURE 3 – Cercle des corrélations de l'AFC sur les tranches d'âges et les habitudes alimentaires

4.1.4 Etude sur les données seulement avec dépressifs et sans dépressifs

Lorsque que nous réalisons le test chi-deux sur le jeu de données sans les individus dépressifs et avec seulement les individus dépressifs, 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 consacrent les étudiants pour leurs études par jour et leurs temps de sommeil. Pour commencer nous pouvons voir avec le tableau 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 montrer 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 disparaître 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 dépressifs

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 disparaît. La p-valeurs est de $\approx 5,8\%$ ce qui est assez proche des valeurs prises 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 dépressifs

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 prenons 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.

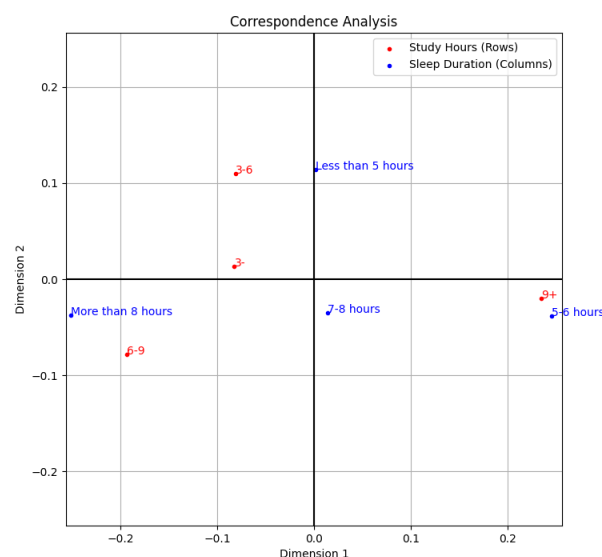


FIGURE 7 – Cercle des corrélations de l'AFC sur le temps d'études et le temps de sommeil

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 correspond à des étudiants passant beaucoup de temps à travailler mais aussi à des étudiants ayant des comportements 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 travaillent 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 élevée 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.

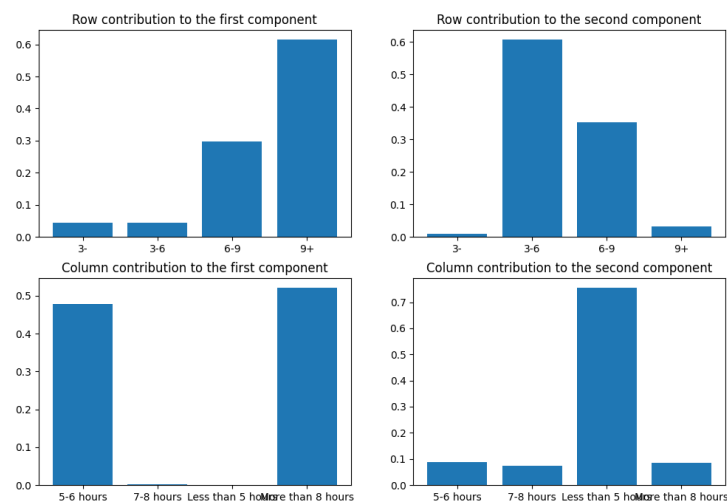


FIGURE 8 – Contribution des catégories de temps d'études et temps de sommeil dans les composantes principales

4.3 Etude du lien entre temps de sommeil et habitudes alimentaires

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

Nous pouvons voir dans le tableau en figure 9 que les effectifs sont globalement équilibrés dans le sens où il y a environ autant de personnes ayant des habitudes alimentaires saines, modérées ou non saines, et de même pour le temps de sommeil. Le test chi-deux nous donne un p-valeur de $\approx 2\%$, ce qui indique que les 2 variables sont significativement corrélées. Ainsi, les personnes ayant un régime alimentaire sain par exemple, pourront avoir tendance à adopter un certain type de sommeil.

Sleep Duration	Dietary Habits		
	Healthy	Moderate	Unhealthy
Less than 5 hours	41	45	37
5-6 hours	52	41	30
7-8 hours	37	44	47
More than 8 hours	31	41	55

FIGURE 9 – Table de contingence entre le temps de sommeil et les habitudes alimentaires

4.3.2 Test chi-deux sur les individus dépressifs et non dépressifs

Après exécution du test chi-deux sur les individus dépressifs, on obtient une p-valeur de $\approx 33\%$, tandis que chez les personnes dépressives, on obtient $\approx 2\%$. Ainsi la dépendance n'est conservée que chez les personnes non dépressives.

Cette indépendance chez les personnes dépressives met en avant l'origine multifactorielle de cette maladie qui perturbe l'ensemble des habitudes de vie. Cela peut conduire à de l'hypersomnie comme de l'insomnie, à une perte d'appétit ou du grignotage ce qui perturbe le lien habituel que l'on observe chez les personnes non dépressives qui ont des comportements alimentaires et des habitudes de sommeil plus stables et cohérents.

La corrélation des non dépressifs pourrait compoenser l'absence de corrélation chez les dépressifs et former la corrélation gloale que l'on observe.

4.3.3 Etude des cercles de corrélation

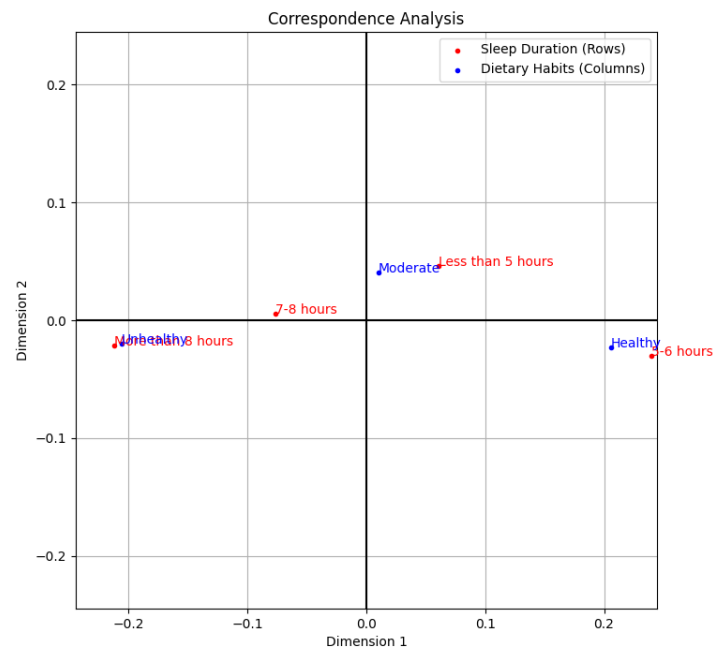


FIGURE 10 – Cercle des corrélations de l'AFC entre le temps de sommeil et les habitudes alimentaires sur l'ensemble des individus

On observe dans la figure 10 un fort lien de corrélation d'une part entre les personnes ayant un temps de sommeil de 5-6 heures et des habitudes alimentaires saines, et d'autre part entre les personnes dormant beaucoup et mangeant de manière peu saine. Les autres points, trop proches du centre, ne sont pas interprétables.

De plus, que ce soit chez les personnes dépressives ou non dépressives, on remarque dans la figure 11 que les tendances semblent persister, à la différence près que chez les dépressifs, les personnes dormant entre 7 et 8 heures tendent plus à avoir un mauvais régime alimentaire, et que chez les non dépressifs, le lien entre le régime sain et 5 à 6 heures de sommeil semble moins certains.

Il serait intéressant de pouvoir observer si ces personnes sont sédentaires ou non. Des personnes sédentaires, qu'elles soient dépressives ou non, pourraient avoir perdu un intérêt pour la cuisine et consacraient leur temps à dormir. Cela apparaît particulièrement cohérent chez les personnes dépressives qui, dans leurs phases de sédentarité, seraient enclines à de l'hypersomnie et des habitudes alimentaires peu saines comme du grignotage ou un manque d'attrait pour la cuisine. Ce comportement n'est pas non plus incohérent chez des personnes non dépressives mais qui présentent cependant une importante sédentarité. Un sommeil long peut être associé à un manque de dynamisme ou à une sédentarité souvent corrélés à un régime alimentaire peu sain. A l'opposé du cercle, le manque de sommeil pourrait être associé à un mode de vie actif pouvant et voulant être compensé par une alimentation saine. Les personnes actives ont les finances et le dynamisme qui permet de mieux subvenir à leurs besoins alimentaires. Le groupe de personnes dépressives peut tout de même présenter des personnes dans la vie active d'où cette corrélation aussi dans leur effectif. Chez les personnes non dépressives, cette corrélation est probablement moins forte car leur état mental leur octroie une plus grande perméabilité dans leur rythme de vie. Elles ne vont pas nécessairement se diriger vers des comportements extrêmes. Les comportements extrêmes sont les plus déterminants dans les corrélations observées.

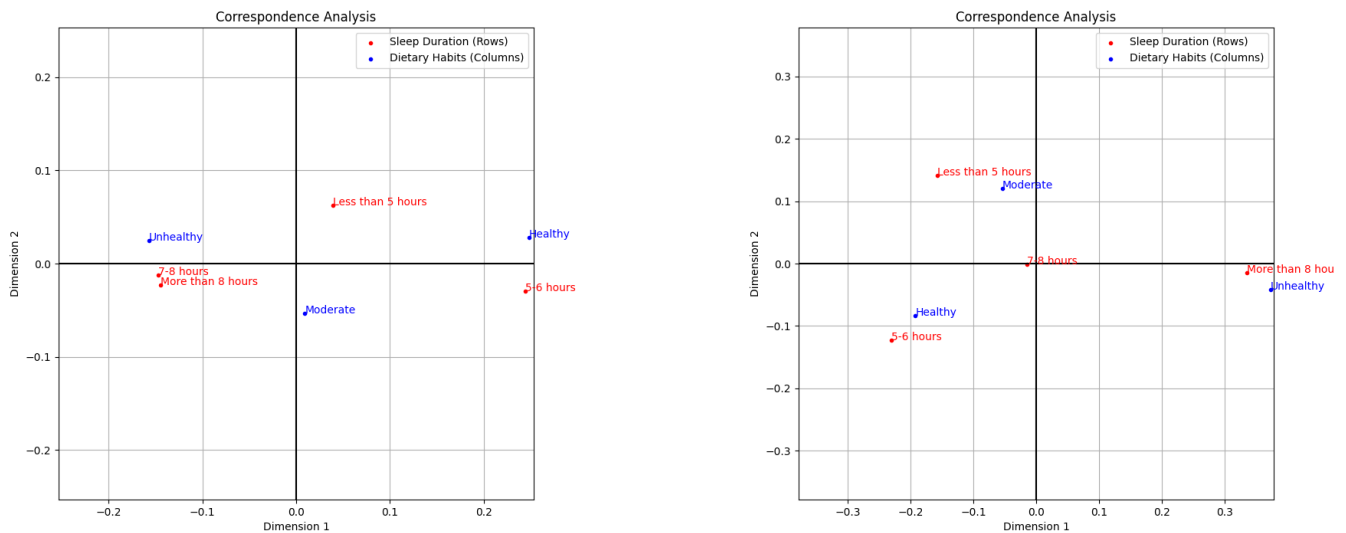


FIGURE 11 – Cercles des corrélations de l'AFC entre le temps de sommeil et les habitudes alimentaires pour les dépressifs (à gauche) et les non dépressifs (à droite)

4.3.4 Contributions des variable dans les composantes principales

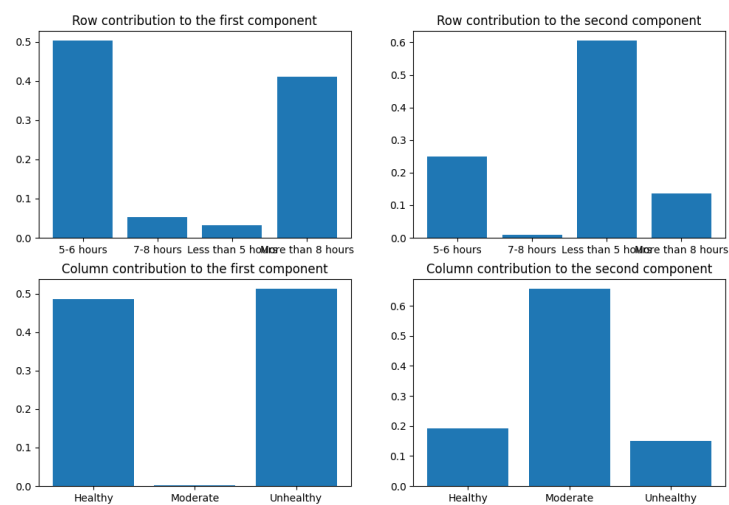


FIGURE 12 – Contributions des catégories de temps de sommeil et de qualité d'alimentation dans composantes principales pour l'ensemble des individus

Sur l'effectif complet de la figure 12 la première composante était bien plus significative que la seconde, sa valeur propre étant 28 fois plus élevée. Elle capte l'essentiel de la variance des données. La seconde variable illustre des comportements plus résiduels. Sur cette première composante, on observe le comportement décrit jusqu'à présent, les comportements extrêmes, pour le sommeil ou les habitudes alimentaires, représentent les contributions les plus importantes. Sur la seconde composante, on observe que les plus grandes contributions sont des comportements que l'on n'a pas réussi à lier à d'autres. Ceci est justifié par la bien moindre valeur de sa valeur propre associée.

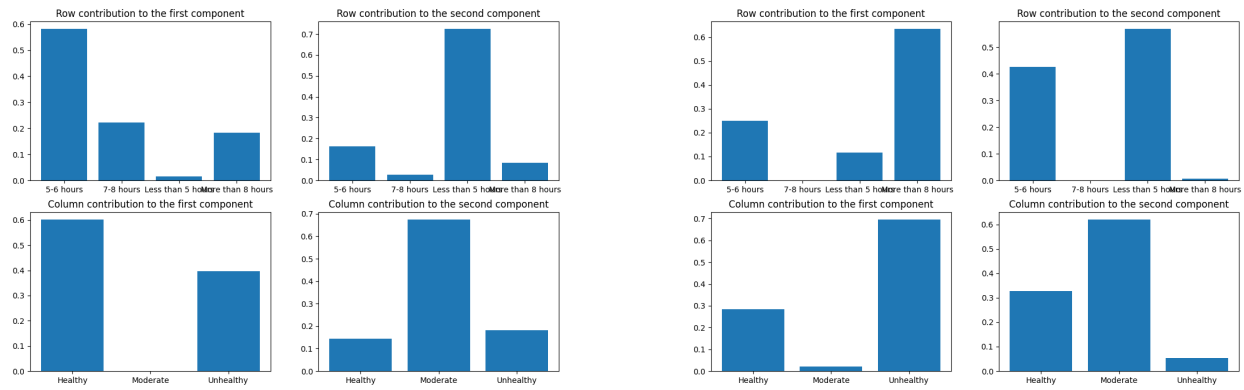


FIGURE 13 – Contributions des catégories de temps de sommeil et de qualité d'alimentation dans composantes principales pour les dépressifs (à gauche) et les non dépressifs (à droite)

Chez les personnes dépressives, selon la figure 13, la première composante, c'est l'association 5-6 heures de sommeil et régime healthy qui sont mises en avant dans les contributions. La seconde composante montre la même tendance que pour le groupe complet, soit une grande contribution de facteurs décorrélés des autres dans notre étude. C'est une observation que nous pouvons aussi faire chez les personnes non dépressives toujours d'après la figure 13. Cependant, pour la première composante, c'est à l'inverse le lien entre unhealthy et 8 heures de sommeil ou plus qui est mis en avant.

De manière générale, les comportements polarisés pourraient être associés à des profils typiques, bien définis. Par exemple, un comportement sédentaire pousse à une grande quantité de sommeil et des habitudes alimentaires malsaines et inversement pour un mode de vie plus actif. Ce sont deux modes de vie bien différenciés mais qui sont observés dans les effectifs dépressifs comme non dépressifs.

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'âge 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>.