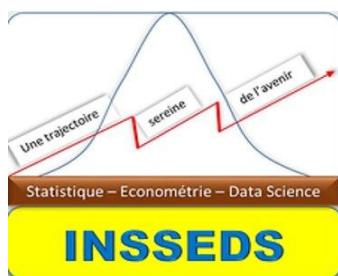


**MINISTÈRE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE****REPUBLIQUE DE COTE D'IVOIRE****INSTITUT SUPERIEUR DES
STATISTIQUES D'ECONOMETRIES ET
DATASCIENCE****UNION-DISCIPLINE-TRAVAIL****MASTER 2
STATISTIQUE-ECONOMETRIE-DATA SCIENCE****MINI-PROJET****INFERENCE STATISTIQUE****ANALYSE DE LA DEPRESSION CHEZ LES ETUDIANTS****ANNEE ACADEMIQUE :****2024 -2025****NOM: KABA****PRENOM: MAHAMOUD TOIB****ENSEIGNANT – ENCADREUR
AKPOSSO DIDIER MARTIAL**

Avant-Propos

Dans un monde en constante évolution, les défis psychologiques, sociaux et économiques auxquels les individus sont confrontés n'ont jamais été aussi prononcés. Parmi ces enjeux, la dépression s'impose comme une problématique de santé publique majeure, touchant toutes les catégories de population, y compris les étudiants. Dans un contexte où les jeunes sont en proie à des pressions académiques, professionnelles et sociales, il devient impératif de comprendre les facteurs qui influencent leur bien-être mental.

Ce travail d'analyse explore des dimensions clés telles que la durée du sommeil, les habitudes alimentaires, le stress financier, et les pensées suicidaires pour établir des liens entre ces variables et la dépression chez les étudiants. Grâce à une approche rigoureuse mêlant statistiques descriptives, tests d'hypothèses et modèles de visualisation, cette étude vise à répondre à des questions essentielles : Quels facteurs contribuent le plus au développement de la dépression ? Quelle est l'interdépendance entre ces facteurs ? Comment peut-on identifier les leviers d'intervention efficaces pour améliorer la santé mentale dans cette population ?

L'objectif est de fournir une analyse approfondie qui n'est pas seulement un diagnostic mais également un outil d'aide à la décision, que ce soit pour les institutions éducatives, les professionnels de la santé, ou les politiques publiques. Ce projet s'inscrit dans une démarche humaniste et scientifique, cherchant à apporter une compréhension plus fine de ce phénomène complexe et multiforme.

Table des matières

Avant-Propos	1
1. Introduction Générale	4
1.1 Contexte et Justification de l'Étude	4
1.2 Problématique	4
1.3 Principaux Résultats Attendus	4
1.4 Méthodologie.....	4
1.4.1 Technique de Prétraitemet Utilisée :	4
1.4.2 Analyses Statistiques et Fondements Théoriques :	5
1.5 . Description du Jeu de Données : Dictionnaire des Données	5
1.6 aperçu du jeu de données.....	6
2. PRETRAITEMENT DES DONNEES	6
2.1 Traitements des valeurs manquantes	6
2.2 TRAITEMENT DES DOUBLONS	8
2.3. IDENTIFICATION ET TRAITEMENTS DES VALEURS EXTREMES	8
1 ère PARTIE : ANALYSE UNIVARIEE DES VARIABLES D'INTERET	9
I. Etudes des variables qualitatives.....	10
II. Etude des variables quantitatives	17
Normalité.....	17
Distribution	19
Résumé numérique	20
2 ème PARTIE : ANALYSE BIVARIEE : Analyse de la liaison de la variable dépression avec les autres variables.....	21
1. Pour une variable qualitative et une quantitative	21
Résumé de l'analyse des variables quantitatives avec dépression	26
3. Pour deux variables qualitatives	26
3 ème PARTIE : ANALYSE DE DONNEES MULTIVARIEES.....	32
1. Méthodes factorielle	32
ACP (Analyse en Composantes Principales).....	32
• ACM (Analyse des Correspondances Multiples)	34
AFC (Analyse Factorielle des Correspondances)	35
Clustering	36
4 -ème parties : réponses aux questions.....	39
Conclusion générale	41
Annexe	42
Tableau de bord	42
Codes utilises	42

Introduction Générale

La dépression est l'un des troubles mentaux les plus répandus au monde, touchant une part croissante de la population, notamment les jeunes étudiants. À travers les pressions académiques, financières, sociales et professionnelles, ces derniers se retrouvent souvent exposés à un ensemble de facteurs de stress qui peuvent avoir des répercussions majeures sur leur santé mentale. Cette étude s'inscrit dans un cadre scientifique visant à mieux comprendre les relations entre ces différents facteurs et la dépression afin d'identifier des leviers d'intervention efficaces.

1.1 Contexte et Justification de l'Étude

Dans le contexte universitaire, les étudiants sont confrontés à des défis variés, allant des attentes académiques élevées aux pressions financières. Ces pressions peuvent exacerber des problèmes de santé mentale, notamment la dépression, qui est souvent accompagnée d'une baisse de performance académique, de pensées suicidaires et d'autres troubles. Par ailleurs, peu d'études ont exploré de manière approfondie les relations entre la dépression, les habitudes de vie (alimentation, sommeil), les pressions externes (financières et académiques) et des variables sociodémographiques. Cette étude se justifie par le besoin urgent de produire des données empiriques pour guider les interventions des décideurs et améliorer la santé mentale des étudiants.

1.2 Problématique

Quels sont les principaux facteurs associés à la dépression chez les étudiants ? Comment des variables comme la durée du sommeil, les habitudes alimentaires, le stress financier et les antécédents familiaux influencent-elles la survenue de la dépression ? À travers cette problématique, nous cherchons à répondre à une question centrale : existe-t-il des différences significatives dans la perception et les effets de ces facteurs entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas ?

1.3 Principaux Résultats Attendus

1. . Identification des principaux facteurs associés à la dépression chez les étudiants, notamment les variables sociodémographiques, les habitudes alimentaires, la durée du sommeil et le stress financier.
1. Mise en évidence des différences significatives dans la satisfaction des études et la satisfaction au travail en fonction de la dépression.
2. Proposition de recommandations basées sur les résultats pour la prévention et la gestion de la dépression dans les milieux universitaires.

1.4 Méthodologie

1.4.1 Technique de Prétraitement Utilisée :

- Nettoyage des Données : Traitement des valeurs manquantes, vérification des doublons et uniformisation des variables catégoriques.
- Codage : Transformation des variables qualitatives en variables numériques (le cas échéant) pour faciliter l'analyse.

- Standardisation : Échelle uniforme pour les variables quantitatives afin de réduire les biais.

1.4.2 Analyses Statistiques et Fondements Théoriques :

- Test de Khi-Deux : Pour évaluer l'indépendance entre la dépression et des variables qualitatives (exemple : habitudes alimentaires, durée du sommeil).
- Tests Non Paramétriques (Mann-Whitney U, Kruskal-Wallis) : Pour comparer les distributions des moyennes ou des médianes de groupes lorsque la normalité des données n'est pas respectée.
- Intervalle de Confiance : Pour estimer les proportions d'étudiants ayant des pensées suicidaires.
- Analyses Descriptives : Moyennes, médianes et écarts-types pour fournir un aperçu général des données.

1.5 . Description du Jeu de Données : Dictionnaire des Données

Nom de la variable	Type	Modalités / Échelle
id	Quantitative (identifiant)	Valeur unique par individu
sexe	Qualitative binaire	Masculin / Féminin
Age	Quantitative continue	Numérique
ville	Qualitative nominale	Plusieurs villes
profession	Qualitative nominale	Étudiant / Autre activité
pression_academique	Quantitative continue (échelle)	Score numérique
pression_liee_au_travail	Quantitative continue (échelle)	Score numérique
moyenne_notes	Quantitative continue	Moyenne sur 20 ou autre
satisfaction_etudes	Ordinal	Échelle de satisfaction
satisfaction_travail	Ordinal	Échelle de satisfaction
duree_sommeil	Qualitative ordinale	Moins de 5h / 5-6h / plus de 6h
habitudes_alimentaires	Qualitative ordinale	Saines / Modérées / Mauvaises
diplome_suivi	Qualitative nominale	BSc / M.Tech / etc.
pensees_suicidaires?	Qualitative binaire	Oui / Non
nombre_heure_travail_etude	Quantitative continue	Nombre d'heures par jour
stress_financier	Quantitative continue (échelle)	Score
antecedents_familiaux_maladies_mentales	Qualitative binaire	Oui / Non
depression	Binaire (cible)	0 = Non, 1 = Oui

1.6 aperçu du jeu de données

	sexe	age	ville	profession	pression_academique	pression_liee_au_travail	moyenne_notes	satisfaction_etudes	satisfaction_travail	duree_sommeil
0	Male	33	Visakhapatnam	Student	5.0	0.0	8.97	2.0	0.0	5-6 hours
1	Female	24	Bangalore	Student	2.0	0.0	5.90	5.0	0.0	5-6 hours
2	Male	31	Srinagar	Student	3.0	0.0	7.03	5.0	0.0	Less than 5 hours
3	Female	28	Varanasi	Student	3.0	0.0	5.59	2.0	0.0	7-8 hours
4	Female	25	Jaipur	Student	4.0	0.0	8.13	3.0	0.0	5-6 hours
	habitudes_alimentaires	diplome_suivi	pensees_suicidaire	nombre_heure_travail_etude	stress_financier	antecedants_familiaux_maladie_mentale	depression			
	Healthy	B.Pharm	Yes	3	1.0	No	Oui			
	Moderate	BSc	No	3	2.0	Yes	Non			
	Healthy	BA	No	9	1.0	Yes	Non			
	Moderate	BCA	Yes	4	5.0	Yes	Oui			
	Moderate	M.Tech	Yes	1	1.0	No	Non			

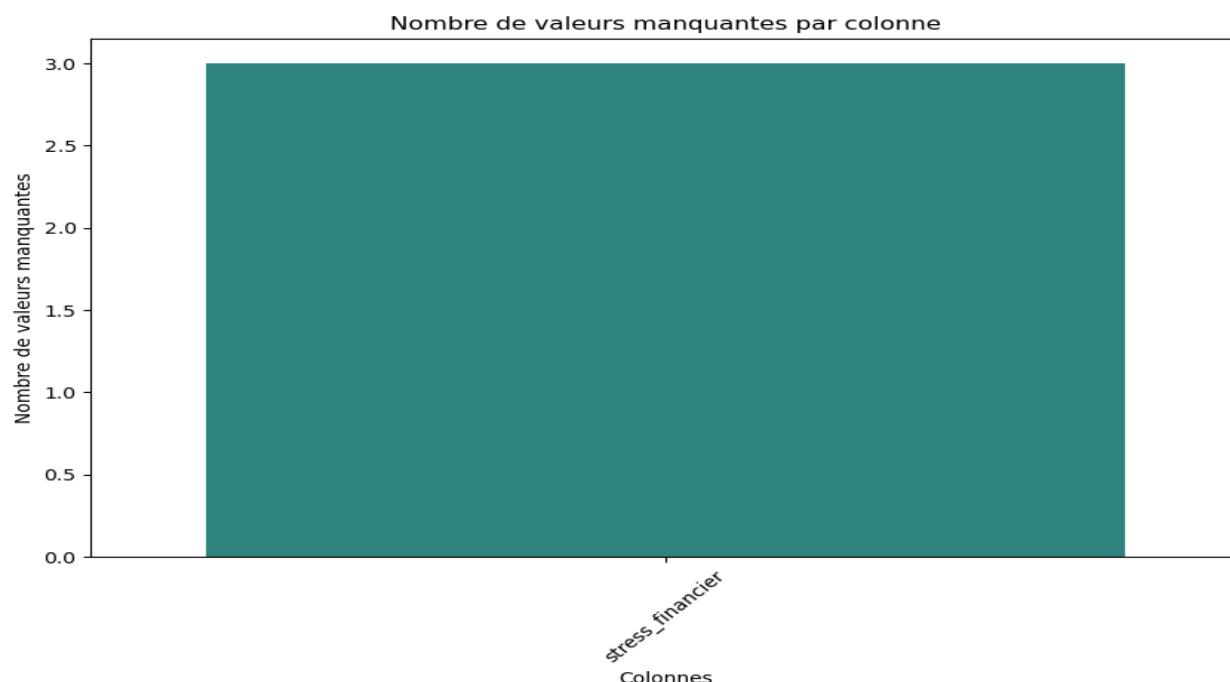
Ce jeu de donnees contient 27901 observations dont 18 variables

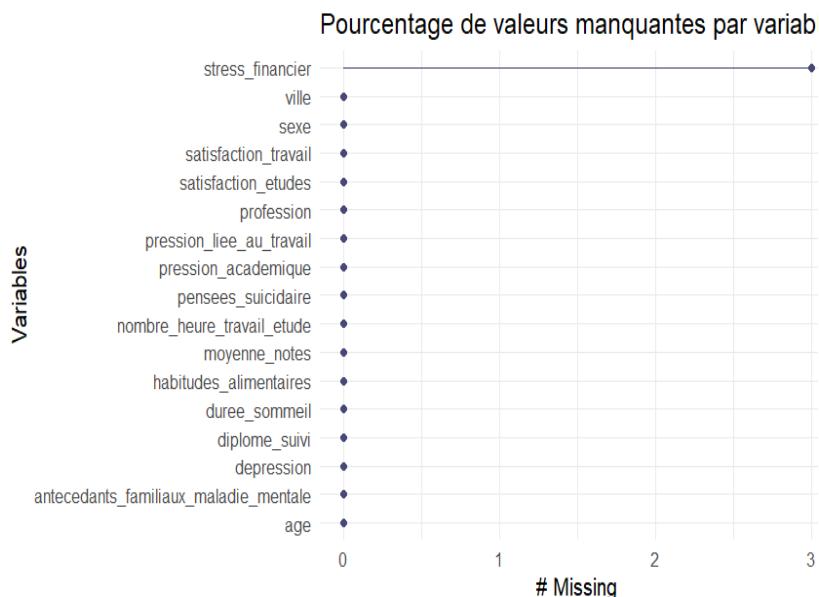
2. PRETRAITEMENT DES DONNEES

2.1 Traitements des valeurs manquantes

Identification des valeurs manquantes dans le jeu de données, calcul de leur fréquence et prise de décisions sur leur suppression ou leur gestion

➤ Visualisation des valeurs manquantes





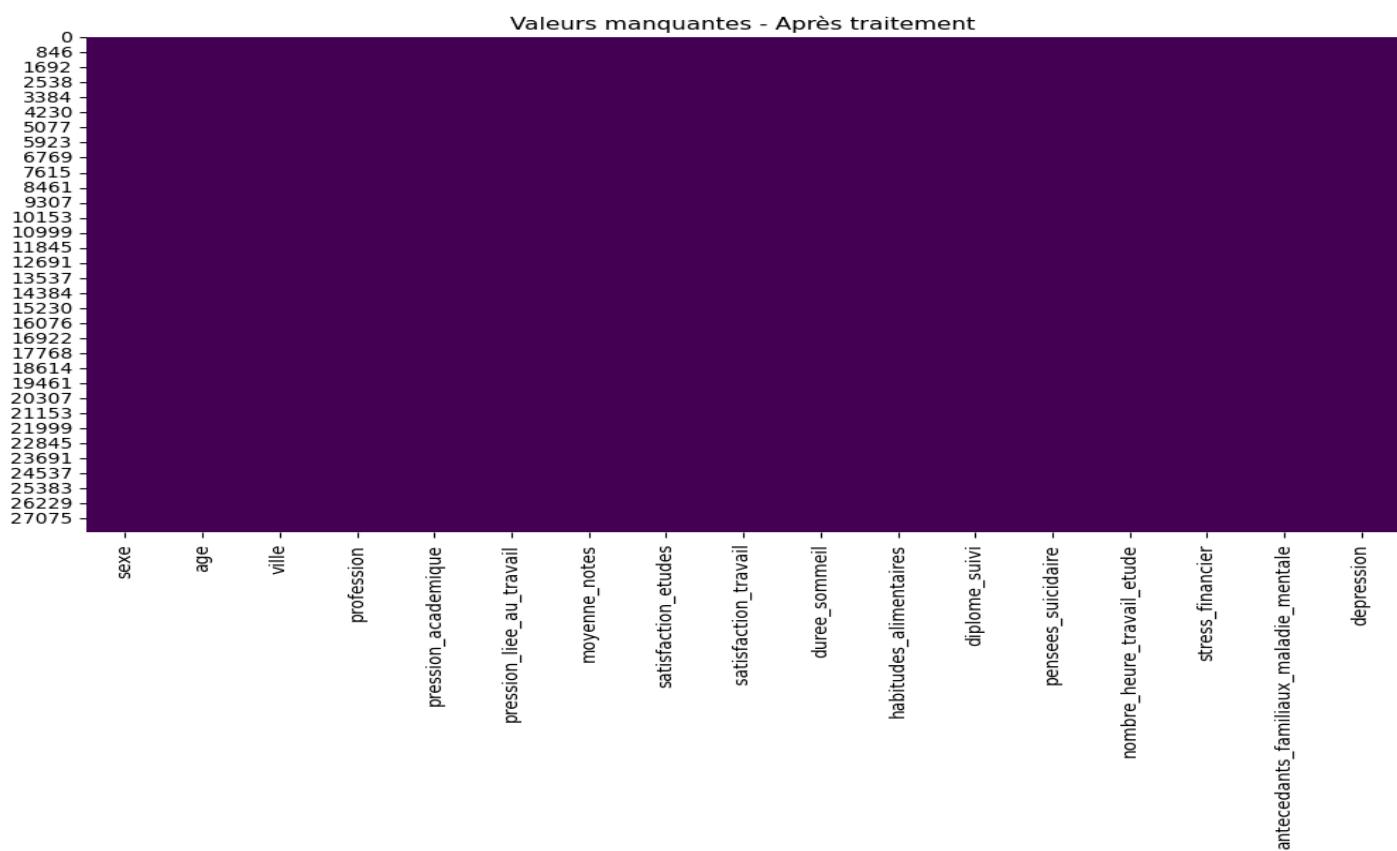
Après la visualisation nous voyons avec exactitude que sur les **27901** observations il n'y a que **3** valeurs manquantes et ces valeurs manquantes proviennent de la variables **stress_financier**. Ces trois valeurs sont tellement insignifiantes sur toutes les observations qu'on arrive à peine à les repérer sur les deux premiers graphiques.

➤ Calcul du pourcentage de valeurs manquantes

Sur les 27901 on a 0.0107523 % de valeurs manquantes

➤ Suppression des individus si moins de 5% sont affectés

Comme sur les 27901 on a 0.0107523 % de valeurs manquantes qui est inférieur à 5% des individus ne disposant pas de valeurs manquantes, alors nous supprimons les individus qui ont des valeurs manquantes. Il nous restera 27898 observations(individus). Nous pouvons constater ci-dessous qu'il n'y a plus de valeurs manquantes.



2.2 TRAITEMENT DES DOUBLONS

Vérification de la présence d'observations répétées et suppression des doublons pour garantir l'intégrité des analyses.

- Détection des observations dupliquées

Après le traitement des doublons avec le langage python, on constate qu'il n'y a pas de doublons

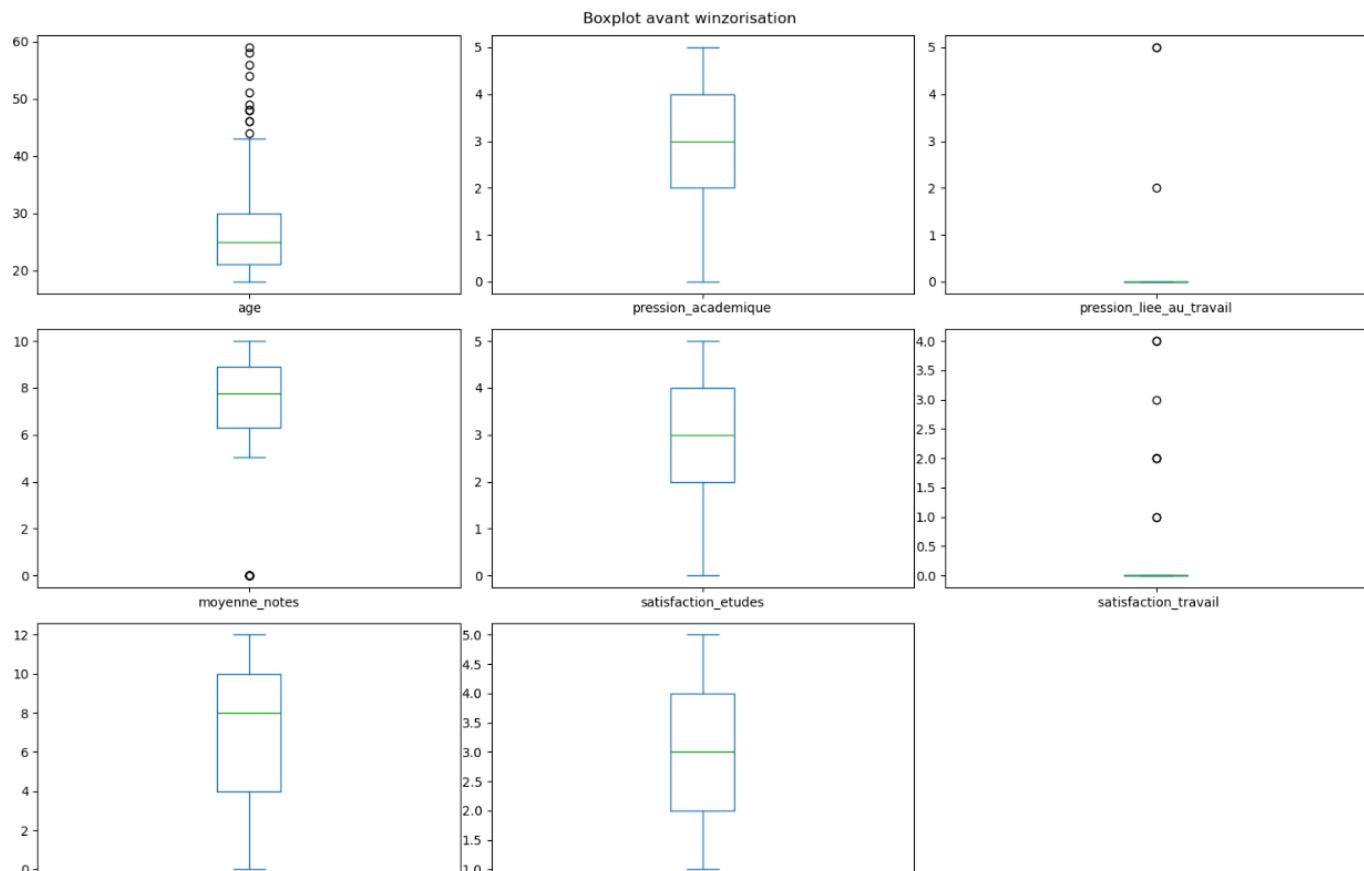
- Suppression des doublons

Vu qu'il n'y a pas de doublons il n'est pas nécessaire de les supprimer.

2.3. IDENTIFICATION ET TRAITEMENTS DES VALEURS EXTREMES

Détection des valeurs aberrantes grâce aux boîtes à moustaches, identification des observations affectées et utilisation de la winzorisation pour limiter leur impact sur l'analyse.

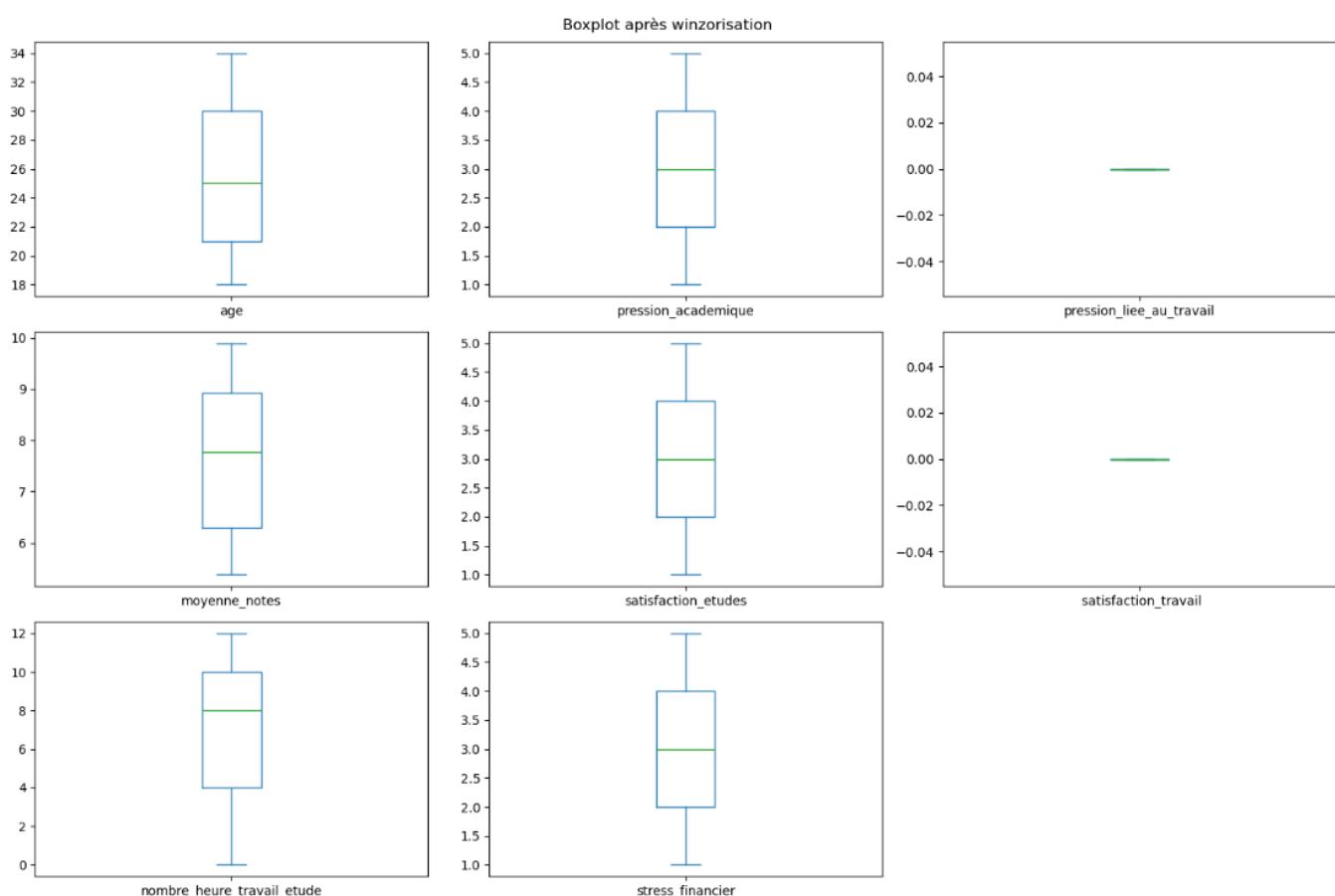
- Détection des valeurs extrêmes avec les boîtes à moustaches des variables quantitatives



- **Graphique 1 correspond à la boite a moustache de l'Age :** nous constatons plusieurs valeurs extrêmes
- **Graphique 2 correspond à la boite a moustache de pression_academique :** absence de valeurs extrêmes

- **Graphique 3 correspond à la boite a moustache de pression_liee_au_travail :** ici on a deux valeurs extrêmes
- **Graphique 4 correspond à la boite a moustache de moyenne_notes presence de valeurs extrême**
- **Graphique 5 correspond à la boite a moustache de satisfaction_etudes :** absence de valeurs extrêmes
- **Graphique 6 correspond à la boite a moustache de satisfaction_travail :** nous constatons plusieurs valeurs extrêmes
- **Graphique 7 correspond à la boite a moustache de nombre_heure_travail_etudes :** absence de valeurs extrêmes
- **Graphique 8 correspond à la boite a moustache de stress_financier :** absence de valeurs extrêmes

➤ **Traitemennt des valeurs extrêmes par winzorisation**



Après la winzorisation, les valeurs extrêmes sont neutraliser

1 ère PARTIE : ANALYSE UNIVARIEE DES VARIABLES D'INTERET

L'analyse univariée est une étape fondamentale dans tout processus d'exploration de données, car elle permet de comprendre les caractéristiques de chaque variable individuellement avant d'examiner leurs relations avec d'autres. Cette première partie, dédiée à l'analyse univariée des variables d'intérêt, se concentre sur l'examen

détaillé de chaque variable pour identifier des tendances, des distributions, et des valeurs anormales ou inattendues.

I. Etudes des variables qualitatives...

L'étude des variables qualitatives consiste à analyser les différentes catégories ou modalités d'une variable afin de comprendre leur répartition et leur influence. Cette analyse est essentielle pour des variables comme le sexe, la profession, les habitudes alimentaires, ou encore la présence de dépression. Elle permet de calculer des fréquences absolues et relatives, de repérer des tendances et d'identifier des catégories dominantes.

Les graphiques en barres, les diagrammes circulaires (camemberts), et les tableaux de contingence sont des outils couramment utilisés pour visualiser et interpréter ces données qualitatives. Cette étape fournit une base solide pour explorer des relations entre variables dans des analyses ultérieures.

- Sexe

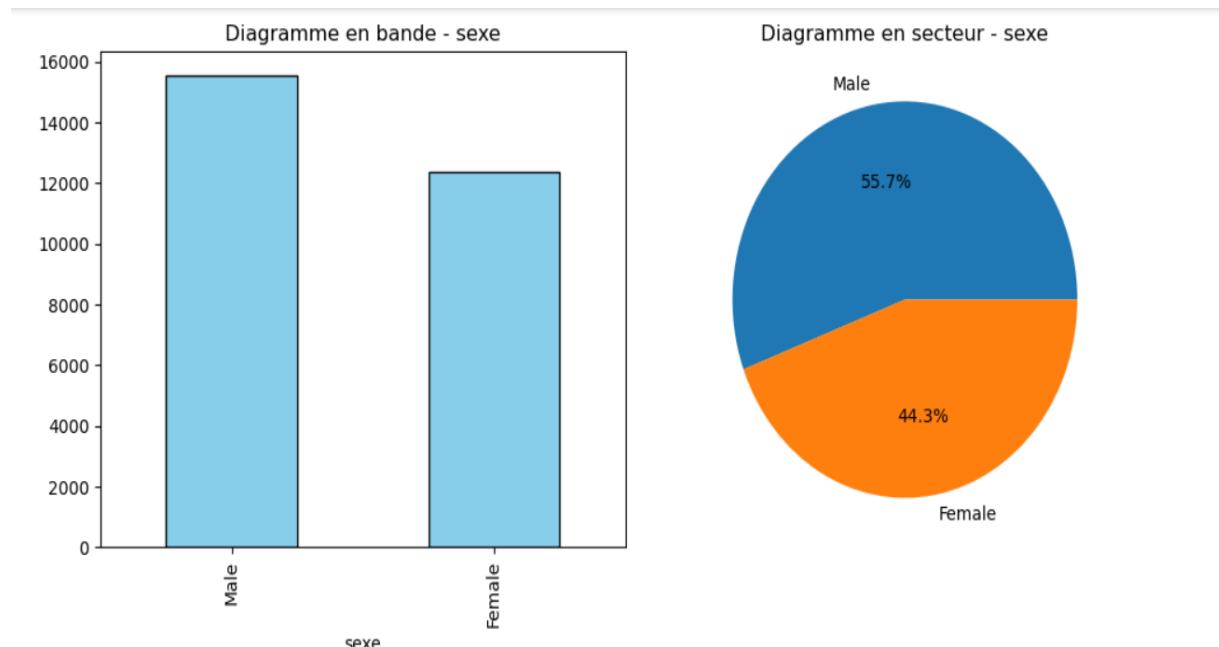


Tableau pour sexe :

Effectif Fréquence (%)		
sexe		
Male	15546	55.72
Female	12352	44.28

Le diagramme en bande montre que la proportion des hommes est légèrement supérieure à celle des femmes dans l'échantillon. Le diagramme en secteur confirme cela, indiquant que 55,7 % des répondants sont des hommes contre 44,3 % de femmes. Cela suggère une distribution déséquilibrée entre les sexes dans les données.

Aussi Nous voyons qu'il y a plus de femmes étudiantes que de d'hommes étudiants soit 15546 étudiantes (55.72%) et

12352 étudiants (44.28%).

- Ville

Diagramme en bande - ville

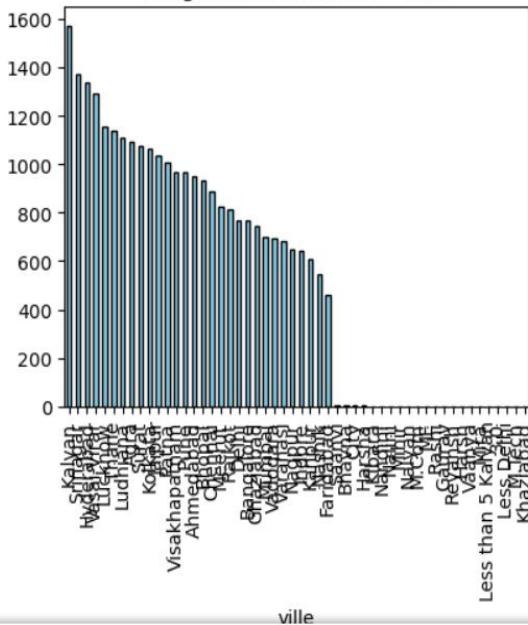
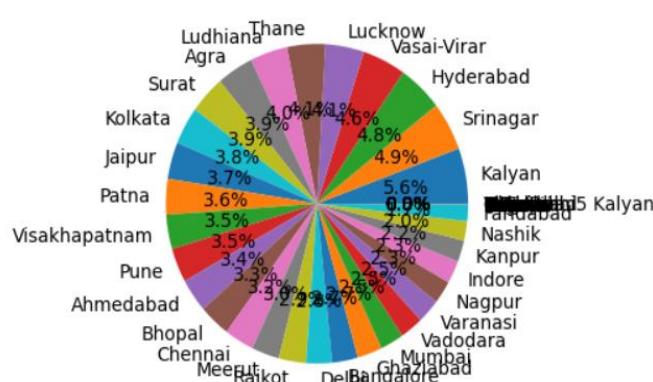


Diagramme en secteur - ville

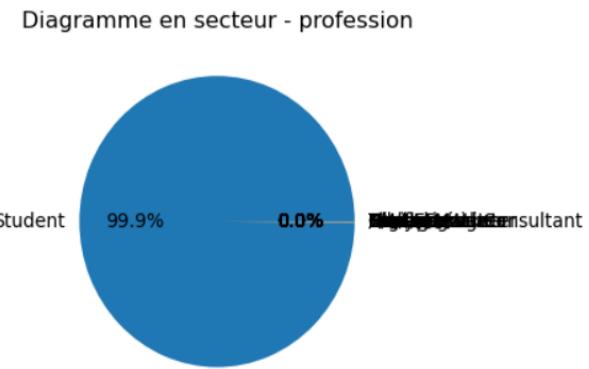
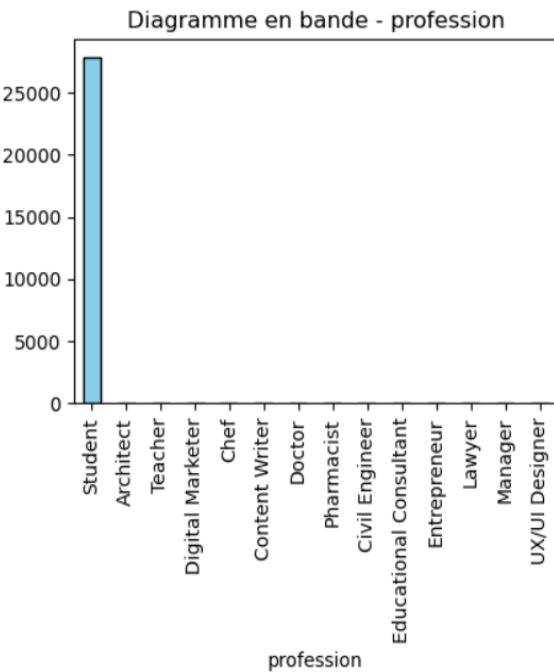


Ville	Effectif	Fréquence (%)
Kalyan	1570	5.63
Srinagar	1372	4.92
Hyderabad	1339	4.80
Vasai-Virar	1290	4.62
Lucknow	1155	4.14
Thane	1139	4.08
Ludhiana	1111	3.98
Agra	1094	3.92
Surat	1078	3.86
Kolkata	1065	3.82
Jaipur	1036	3.71
Patna	1007	3.61
Visakhapatnam	969	3.47
Pune	968	3.47
Ahmedabad	951	3.41
Bhopal	934	3.35
Chennai	885	3.17
Meerut	825	2.96
Rajkot	816	2.92
Delhi	768	2.75

Ici on a les 20 premières ville les plus fréquentes dans le jeu de données

Certaines villes comme **Kalyan**, **Hyderabad**, **Srinagar** regroupent une grande part des étudiants, ce qui peut indiquer une concentration géographique liée à des universités ou conditions spécifiques. Les participants viennent principalement de villes comme Kalyan (5,63 %), Hyderabad (4,80 %), et Srinagar (4,92 %).

- Profession



📊 Diagramme en bande (à gauche) – profession

- Il montre que la profession "Student" (Étudiant) domine très largement toutes les autres professions.
- Les autres professions comme *Architect*, *Teacher*, *Doctor*, etc. ont des effectifs très faibles, presque négligeables.
- Cela se voit dans l'échelle des effectifs : la barre des "Students" dépasse 27 000, tandis que les autres professions sont presque à zéro.

🥧 Diagramme en secteur (à droite) – profession

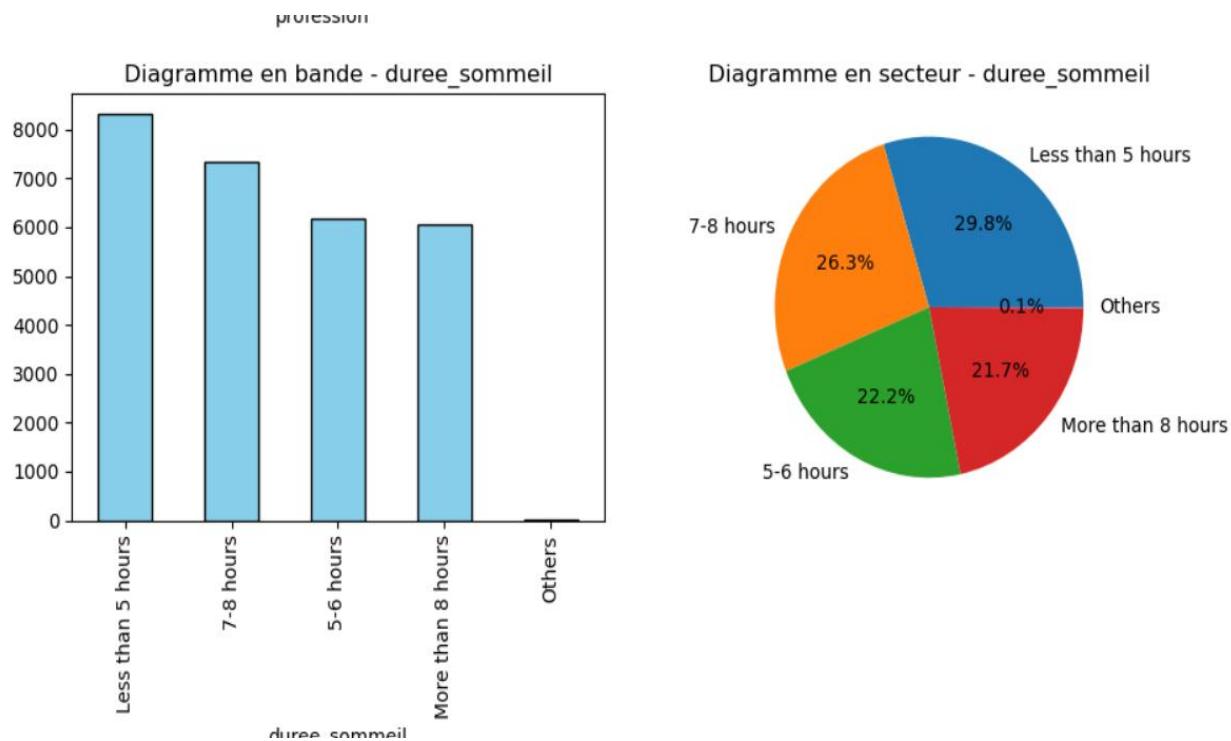
- Le diagramme en secteurs (camembert) confirme cette domination.
- La part des "Students" est de 99,9 %, ce qui signifie que pratiquement tous les individus de l'échantillon sont des étudiants.
- Les autres professions apparaissent en très fines tranches, presque invisibles, avec chacune 0,0 % ou des valeurs très proches de zéro.
- Les étiquettes de ces professions sont superposées et illisibles, ce qui confirme que leur proportion est trop faible pour être représentée clairement dans ce type de graphique.

Profession	Effectif	Fréquence
Étudiant	27867	99.89%
Architecte	8	0.03%
Enseignant	6	0.02%
Digital Marketer	3	0.01%
Chef	2	0.01%
Pharmacien	2	0.01%
Autres	8	0.03%

✓ Interprétation globale

La population étudiée est quasi exclusivement composée d'étudiants. Il y a très peu de représentants d'autres professions. Cela peut indiquer que l'enquête ou l'étude a été réalisée dans un milieu scolaire ou universitaire (ex: une école ou une plateforme

- Durée de sommeil



Durée	Effectif	Fréquence
Moins de 5h	8309	29.78%
5-6h	6181	22.16%
7-8h	7346	26.33%
Plus de 8h	6044	21.66%

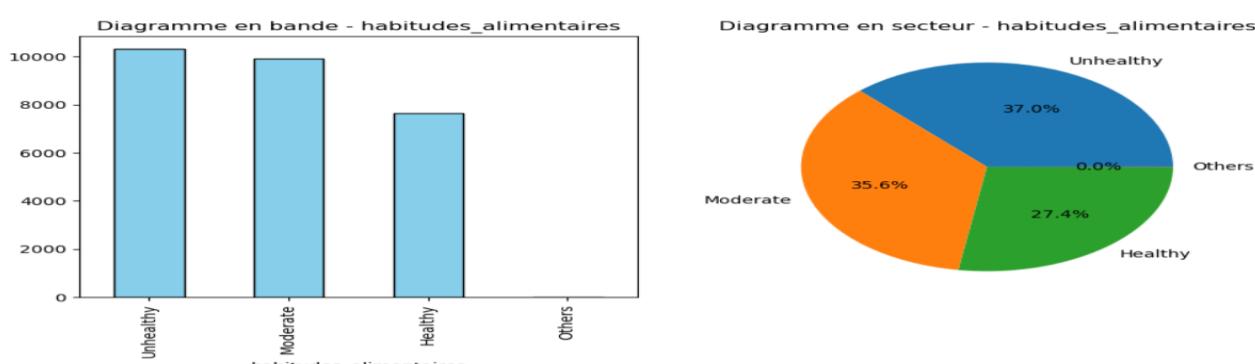
les graphiques illustrent la durée de sommeil des individus de l'échantillon.

• Le **diagramme en barres** montre que les durées de sommeil les plus fréquentes sont **moins de 5 heures** (environ 8000 personnes), suivies de **7-8 heures** (environ 7000). Les catégories **5-6 heures** et **plus de 8 heures** comptent chacune environ 6000 individus, tandis que la catégorie "autres" est marginale avec environ 5000 personnes.

- Le **diagramme en secteur** reflète ces proportions sous forme de pourcentages. La catégorie **moins de 5 heures** domine avec **29,8 %**, tandis que **7-8 heures** et **5-6 heures** représentent **26,3 %** et **22,2 %**, respectivement. Les durées **plus de 8 heures** sont légèrement inférieures à celles des 5-6 heures, avec **21,7 %**.

Cela révèle une prédominance d'individus avec des durées de sommeil relativement courtes (<5 heures), ce qui pourrait être un facteur associé au bien-être ou à la santé mentale

- Habitudes alimentaires



Type d'alimentation	Effectif	Fréquence
Sain	7649	27.42%
Modéré	9921	35.56%
Mauvais	10316	36.98%
Autres	12	0.04%

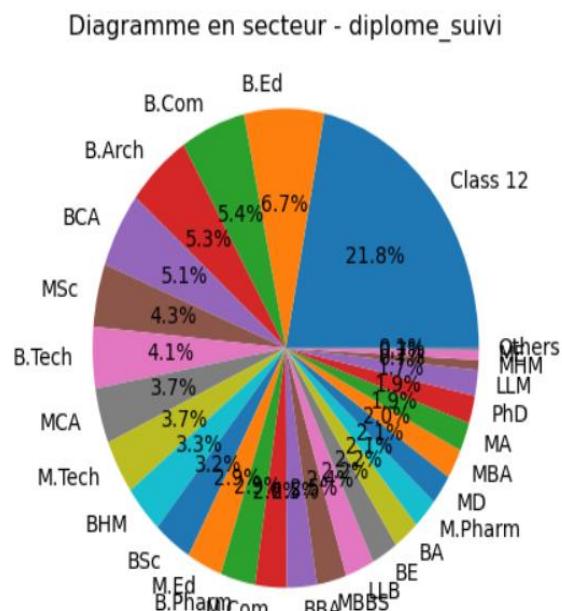
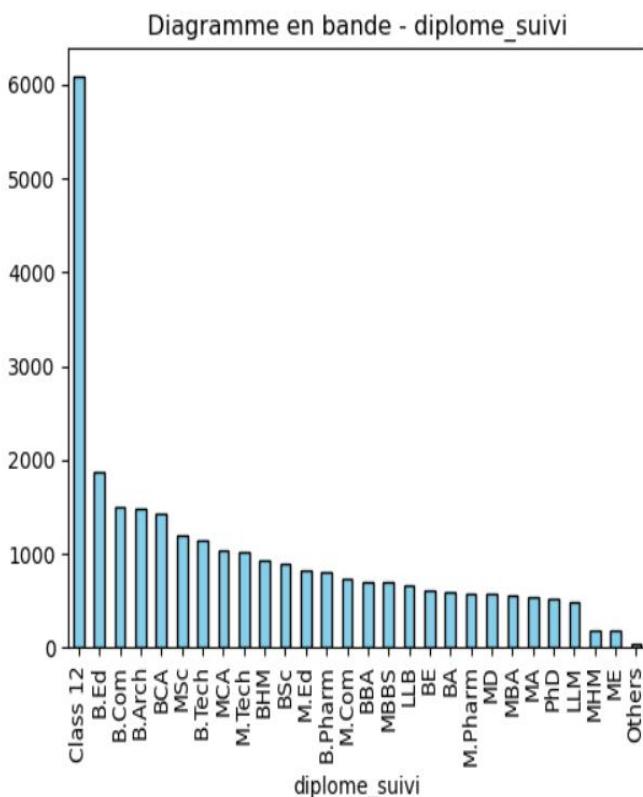
- "Others" sont quasiment absentes.

Diagramme en secteur (droite) :

- **37 %** des individus ont des habitudes malsaines.
 - **35,6 %** ont des habitudes modérées.
 - **27,4 %** ont une alimentation saine.
 - Les **autres types** sont négligeables (**0,0 %**).

Conclusion : La majorité des personnes ont une alimentation non optimale (malsaine ou modérée). L'alimentation saine reste minoritaire dans l'échantillon.

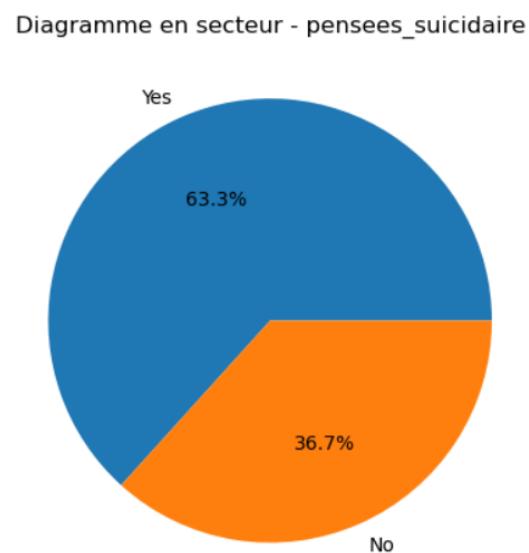
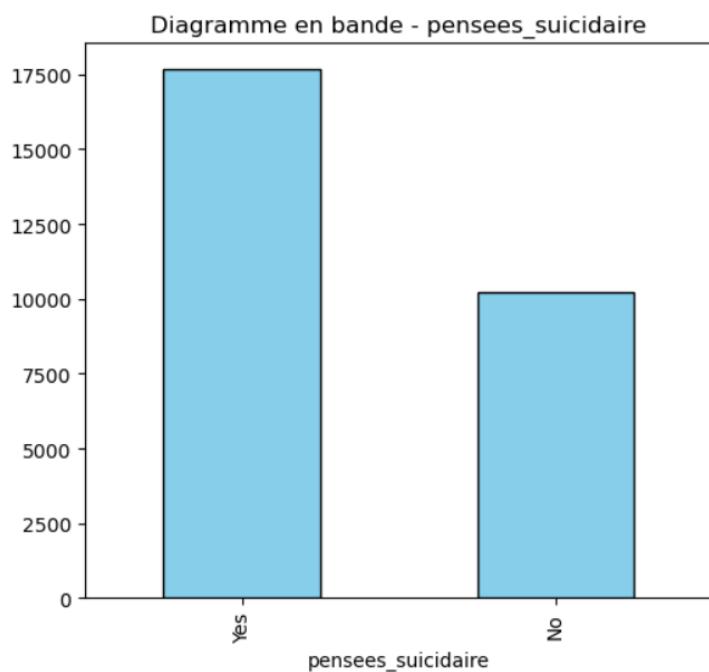
- #### ■ Diplôme suivi



Diplôme suivi	Effectif	Fréquence (%)
Class 12	6 080	21.79
B.Ed	1 866	6.69
B.Com	1 506	5.40
B.Arch	1 478	5.30
BCA	1 432	5.13
MSc	1 190	4.27
B.Tech	1 152	4.13
MCA	1 044	3.74
M.Tech	1 022	3.66
BHM	925	3.32
BSc	888	3.18
M.Ed	821	2.94
B.Pharm	810	2.90
M.Com	734	2.63
BBA	696	2.49
MBBS	695	2.49
LLB	671	2.41
BE	613	2.20
BA	600	2.15
M.Pharm	582	2.09
MD	572	2.05
MBA	562	2.01
MA	544	1.95
PhD	522	1.87
LLM	482	1.73
MHM	191	0.68
ME	185	0.66
Others	35	0.13

Les graphiques montrent la répartition des diplômes suivis par les étudiants. Le diplôme **Class 12** est le plus fréquent avec **21.8 %**, suivi de **B.Com** (6.7 %) et **B.Ed** (5.4 %). Cela suggère une concentration significative d'étudiants dans certaines qualifications académiques, avec une répartition plus faible dans les autres diplômes.

- Pensee suicidaire



Pensées suicidaires	Effectif	Fréquence (%)
Yes	17 656	63.29
No	10 242	36.71

Les graphiques montrent la répartition des pensées suicidaires dans l'échantillon.

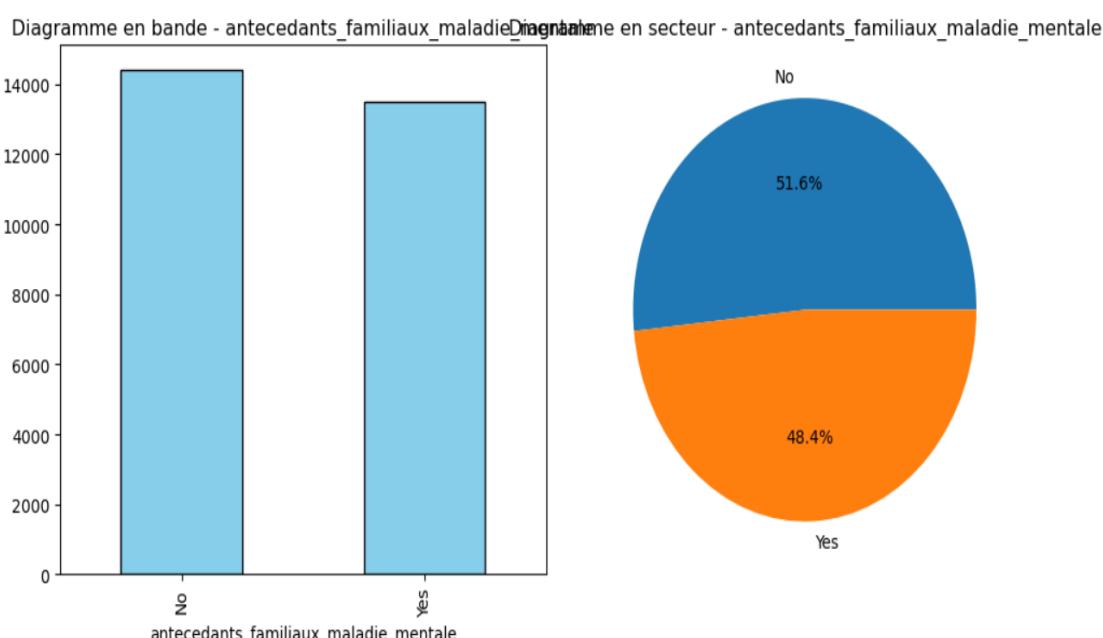
- Le **diagramme en barres** indique que 63,3 % des individus déclarent avoir eu des pensées suicidaires,

contre 36,7 % qui n'en ont pas eu.

- Le **diagramme en secteur** confirme cette majorité en représentant visuellement la prédominance des réponses "Yes" (pensées suicidaires) dans l'ensemble des données.

Ces résultats suggèrent une proportion inquiétante d'individus affectés, nécessitant des interventions ciblées.

■ Antécédents Familiaux



Antécédents familiaux de maladie mentale	Effectif	Fréquence (%)
No	14 397	51.61
Yes	13 501	48.39

Les graphiques montrent la répartition des antécédents familiaux de maladies mentales dans l'échantillon :

- Le **diagramme en barres**

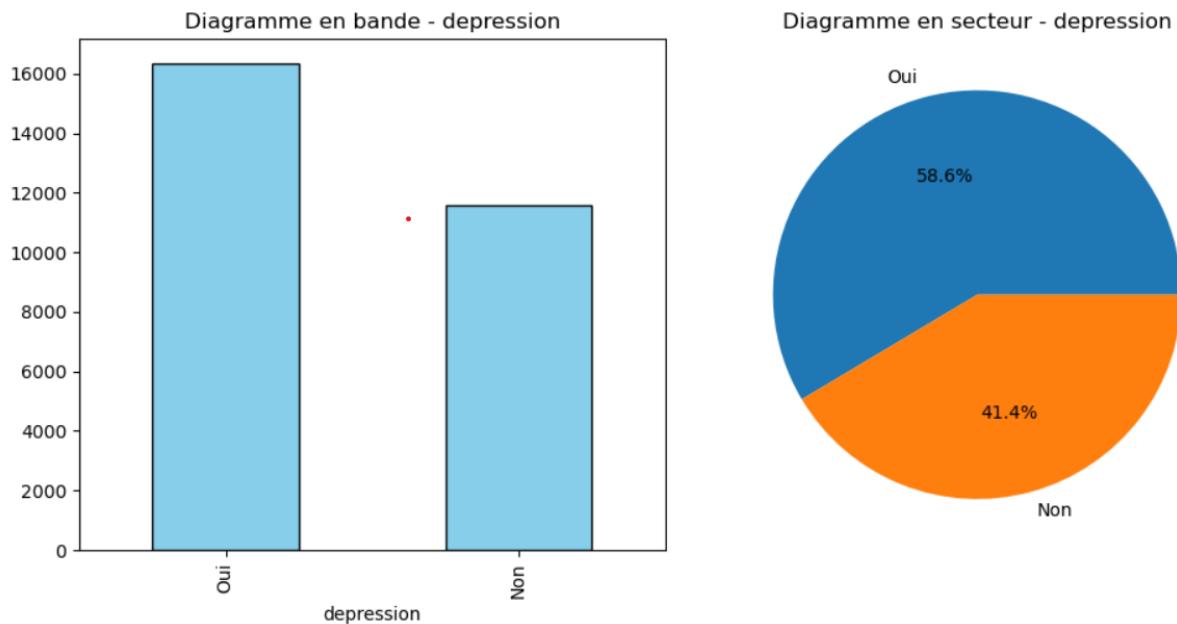
révèle que les réponses sont presque équilibrées entre "0" (pas d'antécédents) et "1" (présence d'antécédents), avec environ **14,000** pour "0" et **13,000** pour "1".

- Le **diagramme en secteur** confirme cette distribution. Il montre que **51.6 %** des répondants n'ont pas d'antécédents familiaux de maladies mentales, tandis que **48.4 %** en ont.

Ces résultats suggèrent une proportion significative d'individus avec des antécédents familiaux, ce qui peut potentiellement influencer leur santé mentale, notamment en lien avec la dépression

■ Dépression

Dépression	Effectif	Fréquence (%)
Oui	16 335	58.55
Non	11 563	41.45



les graphiques présentent la répartition de la dépression au sein de l'échantillon :

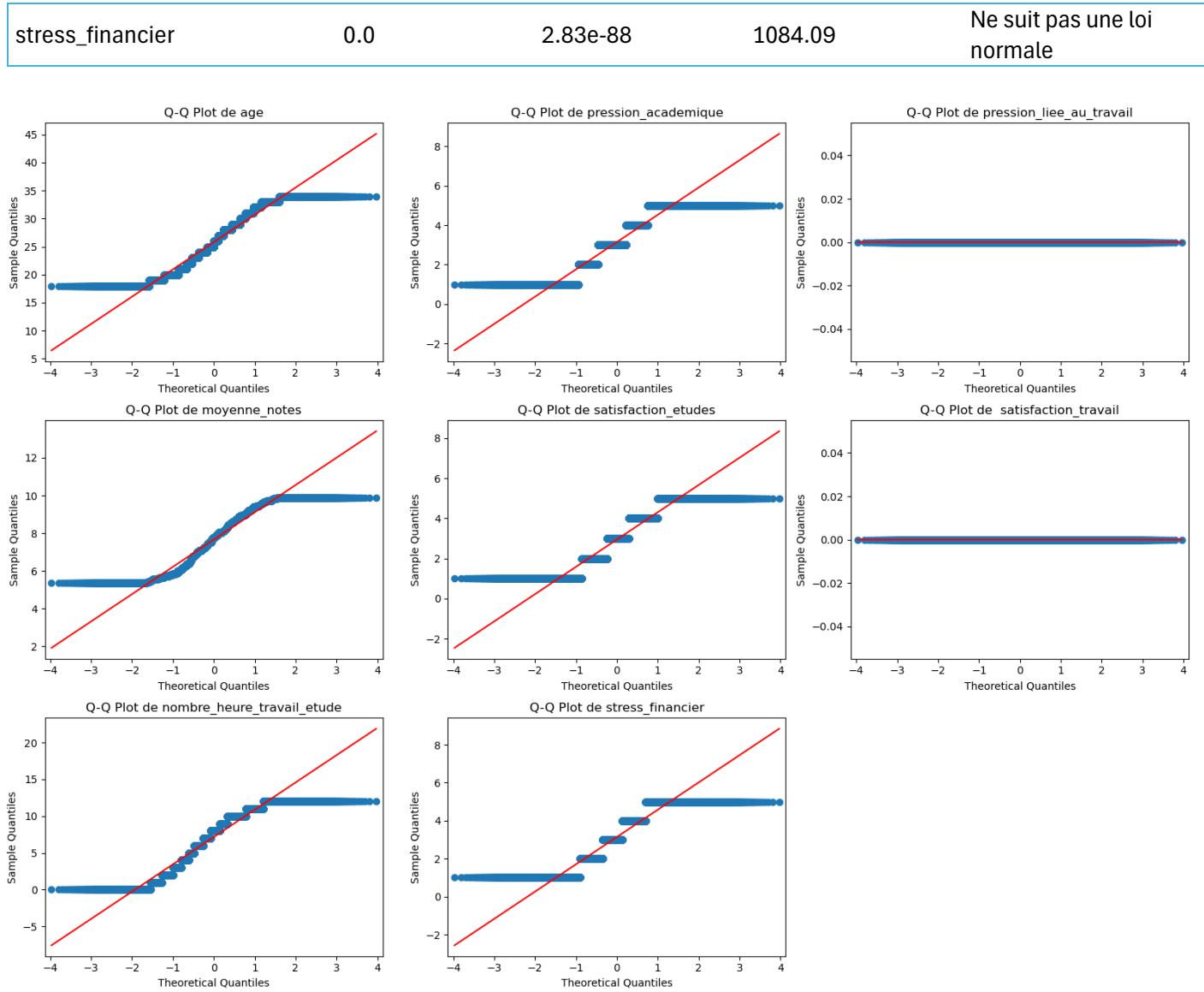
- **Diagramme en barres** : Il montre que **58,6 %** des individus souffrent de dépression ("Oui"), tandis que **41,4 %** n'en souffrent pas ("Non"). Cela reflète une prédominance des cas de dépression parmi les participants.
- **Diagramme en secteur** : Il illustre ces proportions sous forme de pourcentages visuels, confirmant la majorité significative des individus dépressifs.

Ces résultats mettent en évidence l'importance de la problématique de la dépression dans l'échantillon étudié

II. Etude des variables quantitatives

Normalité

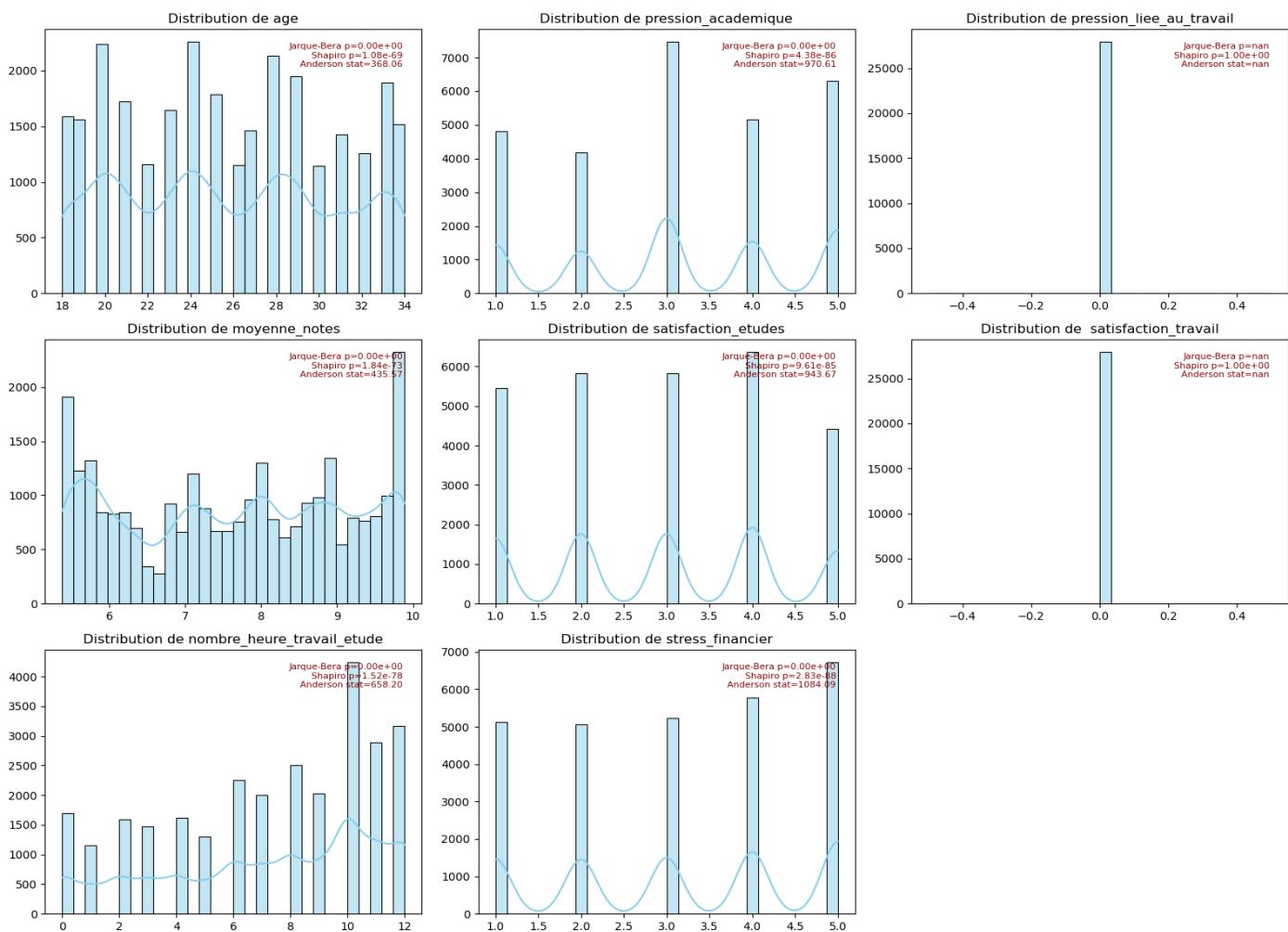
Variable	Jarque-Bera p-value	Shapiro-Wilk p-value	Anderson-Darling stat	Interprétation
age	0.0	1.08e-69	368.06	Ne suit pas une loi normale
pression_academique	0.0	4.38e-86	970.61	Ne suit pas une loi normale
pression_liee_au_travail	NaN	1.0	NaN	Suit une loi normale (selon SW)
moyenne_notes	0.0	1.84e-73	435.57	Ne suit pas une loi normale
satisfaction_etudes	0.0	9.61e-85	943.67	Ne suit pas une loi normale
satisfaction_travail	NaN	1.0	NaN	Suit une loi normale (selon SW)
nombre_heure_travail_etude	0.0	1.52e-78	658.20	Ne suit pas une loi normale



- Variables avec déviation modérée** : Les variables comme **age**, **pression_academique**, et **moyenne_notes** présentent des déviations modérées par rapport à la ligne rouge, ce qui indique qu'elles ne suivent pas une distribution parfaitement normale.
- Variables très proches de la normalité** : **satisfaction_etudes** et **nombre_heure_travail_etude** montrent une meilleure alignment avec la ligne rouge, suggérant qu'elles pourraient être plus proches d'une distribution normale.
- Variables sans variation** : Les Q-Q plots pour **pression_liee_au_travail** et **satisfaction_travail** révèlent des points plat le long de l'axe des x, indiquant un manque de variation dans les données (valeurs constantes ou presque).
- Stress financier** : Une légère déviation est visible, ce qui peut refléter une asymétrie ou une concentration importante autour d'une valeur spécifique.

Ces Q-Q plots permettent d'évaluer rapidement la normalité des données, ce qui est utile pour choisir des tests statistiques appropriés pour les analyses futures.

Distribution



Les graphiques présentent les **distributions des variables quantitatives**, où chaque histogramme montre les fréquences des valeurs spécifiques avec une ligne de densité superposée :

- Age** : L'âge semble être réparti de manière asymétrique, avec une concentration autour d'une tranche spécifique. Cela pourrait refléter une population étudiante relativement homogène.
- Pression académique et pression liée au travail** : Ces variables montrent des distributions légèrement asymétriques, où beaucoup d'individus déclarent un niveau de pression moyen à élevé.
- Moyenne des notes** : Les données semblent concentrées autour d'une moyenne typique, avec quelques valeurs extrêmes potentiellement significatives.
- Satisfaction des études et satisfaction du travail** : Les distributions sont plus uniformes, indiquant une variation significative dans les niveaux de satisfaction.
- Nombre d'heures de travail ou d'étude** : La majorité semble consacrer un nombre moyen d'heures, avec des individus travaillant ou étudiant intensément en dehors de cette plage.
- Stress financier** : Cette variable montre une concentration autour des niveaux moyens à élevés, ce qui pourrait indiquer un défi courant pour la population étudiée.

Ces distributions permettent de dégager des tendances importantes dans les données et peuvent éclairer les analyses futures, notamment pour les relations entre ces variables et la dépression ou le bien-être général.

Résumé numérique

Variable	Moyenne	Médiane	Min	Max	Écart-type	Variance	Asymétrie (Skew)	Kurtosis
age	25.81	25.00	18.00	34.00	4.87	23.72	0.07	-1.18
pression_academique	3.14	3.00	1.00	5.00	1.38	1.91	-0.13	-1.16
pression_liée_au_travail	0.00	0.00	0.00	0.00	0.00	0.00	NaN	NaN
moyenne_notes	7.66	7.77	5.38	9.89	1.45	2.09	-0.05	-1.27
satisfaction_etudes	2.94	3.00	1.00	5.00	1.36	1.85	0.01	-1.23
satisfaction_travail	0.00	0.00	0.00	0.00	0.00	0.00	NaN	NaN
nombre_heure_travail_etude	7.16	8.00	0.00	12.00	3.71	13.75	-0.45	-1.00
stress_financier	3.14	3.00	1.00	5.00	1.44	2.07	-0.13	-1.32

Interprétation des statistiques descriptives

L'analyse des variables quantitatives permet de mieux comprendre le profil des étudiants de notre étude :

- **Âge** : La moyenne est de 25,8 ans, avec une faible asymétrie ($skew \approx 0$), ce qui indique une distribution relativement **symétrique**, bien que légèrement aplatie ($kurtosis < 0$). Cela reflète une population assez homogène en âge.
- **Pression académique et stress financier** : Les moyennes tournent autour de 3/5 avec une asymétrie faible, traduisant un **niveau de pression modéré** chez les étudiants. Toutefois, la forme aplatie des distributions indique une **répartition relativement uniforme**, suggérant que les étudiants ressentent cette pression à des niveaux variés.
- **Moyenne des notes** : Moyenne de 7,6/10, relativement bien répartie. La distribution est **presque normale**, ce qui est rassurant sur le plan académique. Cela indique que la plupart des étudiants ont un niveau scolaire satisfaisant.
- **Satisfaction vis-à-vis des études** : Moyenne de 2,94/5. Légère symétrie et kurtosis négatif montrent une **répartition assez homogène**, mais une satisfaction globalement **moyenne**, ce qui peut avoir un lien avec la dépression.
- **Nombre d'heures de travail par semaine** : Moyenne de 7 heures, mais une **asymétrie négative** montre que certains étudiants travaillent très peu tandis que d'autres cumulent étude et travail. Cette hétérogénéité peut être un facteur de stress ou de surcharge.
- **Pression liée au travail et satisfaction au travail** : Ces variables sont nulles chez tous les répondants, ce qui indique que la **majorité des étudiants n'ont pas d'activité professionnelle**. Cela pourrait influencer le stress financier ou la perception de charge de travail.

Conclusion pour le projet

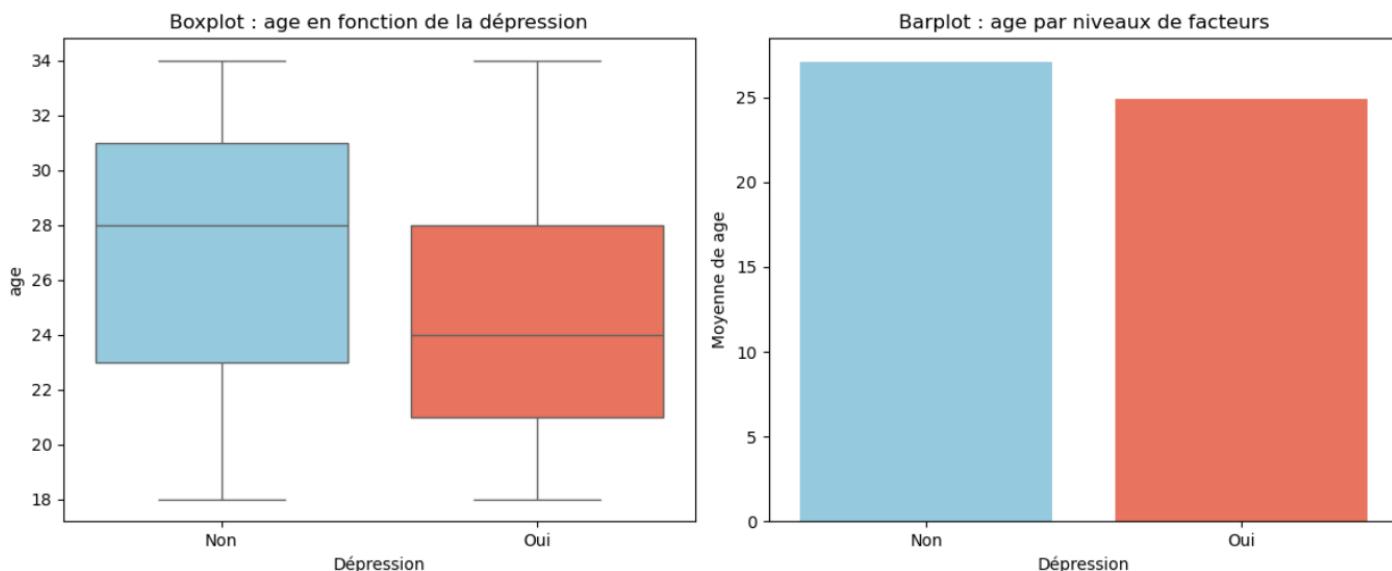
Ces résultats montrent que certaines variables comme **le stress financier, la pression académique et la satisfaction dans les études** varient fortement entre étudiants et pourraient **expliquer ou prédire des épisodes de dépression**. L'absence de normalité dans la majorité des variables implique qu'il faudra **utiliser des tests statistiques non paramétriques** ou des **modèles robustes** dans la suite du projet.

2 ème PARTIE : ANALYSE BIVARIEE : Analyse de la liaison de la variable dépression avec les autres variables...

1. Pour une variable qualitative et une quantitative

- Analyse de la relation entre l'âge et la dépression**

L'âge, en tant que variable quantitative, ne suit pas une distribution normale selon les tests de normalité effectués (par exemple, Shapiro-Wilk ou Jarque-Bera). Par conséquent, les comparaisons entre les groupes (dépression : Oui/Non) ont été réalisées à l'aide du **test non paramétrique de Kruskal-Wallis**, adapté pour les données qui ne respectent pas l'hypothèse de normalité.

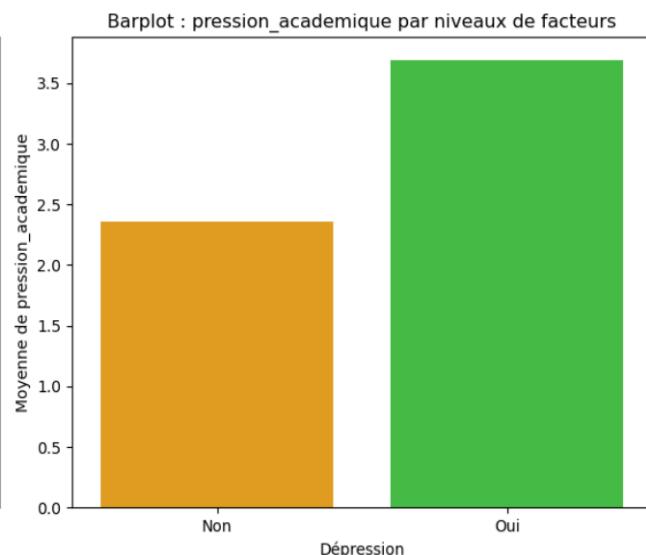
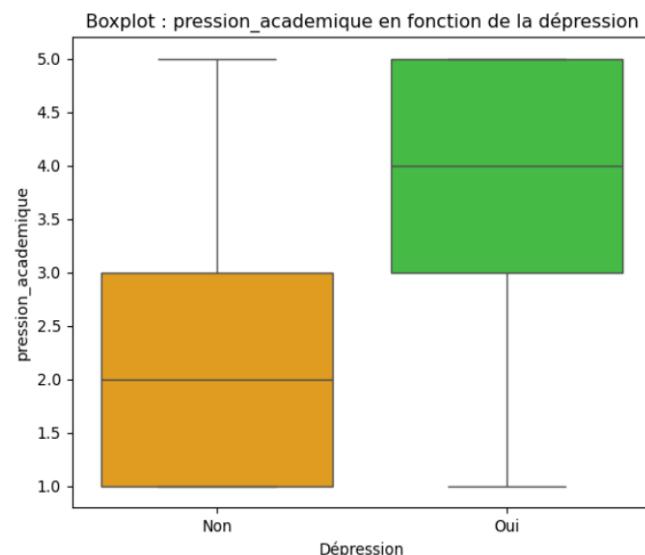


Le **boxplot** montre que la distribution des âges diffère entre les individus avec et sans dépression. Les âges des individus sans dépression sont globalement plus élevés et présentent une plus grande dispersion. Le **barplot**, quant à lui, indique que la moyenne d'âge des individus sans dépression est légèrement supérieure à celle des individus avec dépression.

Le test de **Kruskal-Wallis** confirme cette différence avec une statistique élevée soit 1417.258 et une p-value de 0.0000, indiquant une relation significative entre l'âge et la dépression. Cela suggère que l'âge influence la probabilité de dépression de manière notable.

- Analyse de la relation entre pression académique et la dépression**

La pression académique ne suit pas une loi normale alors pour le test on utilise le test de kruskal (test non paramétrique}

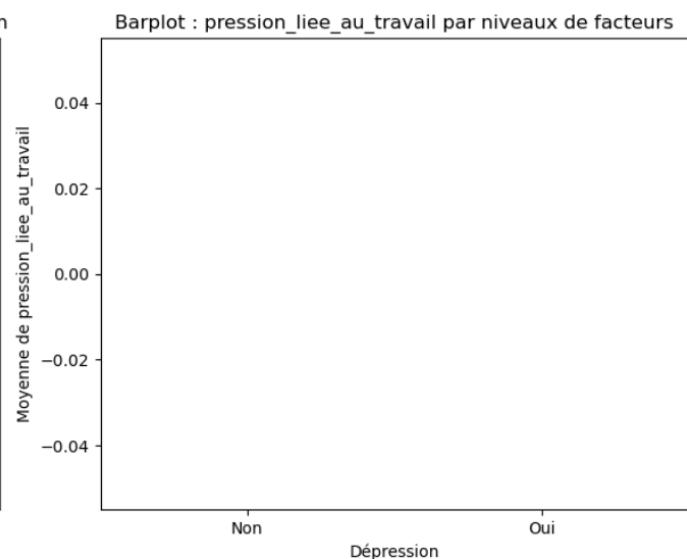
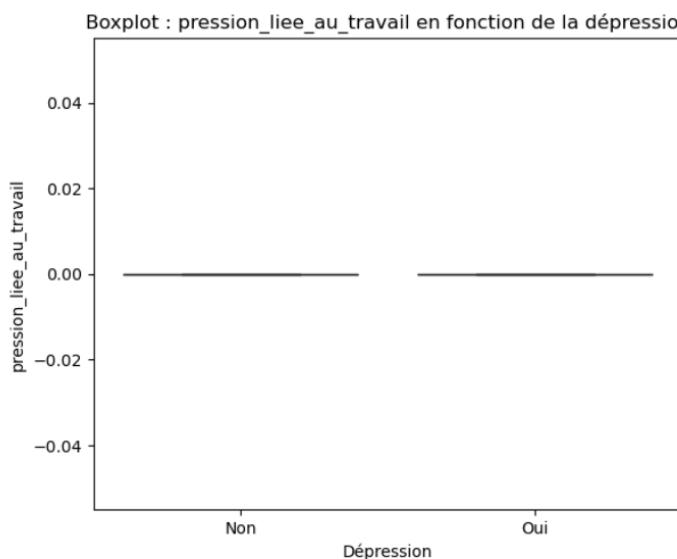


L'analyse de la relation entre la **pression académique** et la dépression révèle des résultats significatifs.

Le **boxplot** montre que les individus souffrant de dépression (Oui) ont une **pression académique plus élevée**, avec des valeurs globalement supérieures et une dispersion notable comparée à ceux qui ne souffrent pas de dépression (Non). Le **barplot**, quant à lui, confirme que la moyenne de pression académique est plus importante chez les personnes avec dépression.

Le test statistique de **Kruskal-Wallis**, avec une statistique extrêmement élevée (**6224.990**) et une **p-value de 0.0000**, indique une **relation significative** entre la pression académique et la dépression. Cela signifie que la pression académique varie nettement selon les niveaux de dépression, suggérant qu'elle pourrait être un facteur explicatif important.

- Analyse de la relation entre pression liée au travail et la dépression**



La **pression liée au travail** présente une moyenne très proche de **0** dans les deux groupes, que ce soit pour les individus avec ou sans dépression, comme indiqué dans le boxplot et le barplot. Cela signifie que cette variable

ne montre pas de variation significative entre les groupes, ce qui peut indiquer qu'elle n'est pas un facteur explicatif important de la dépression.

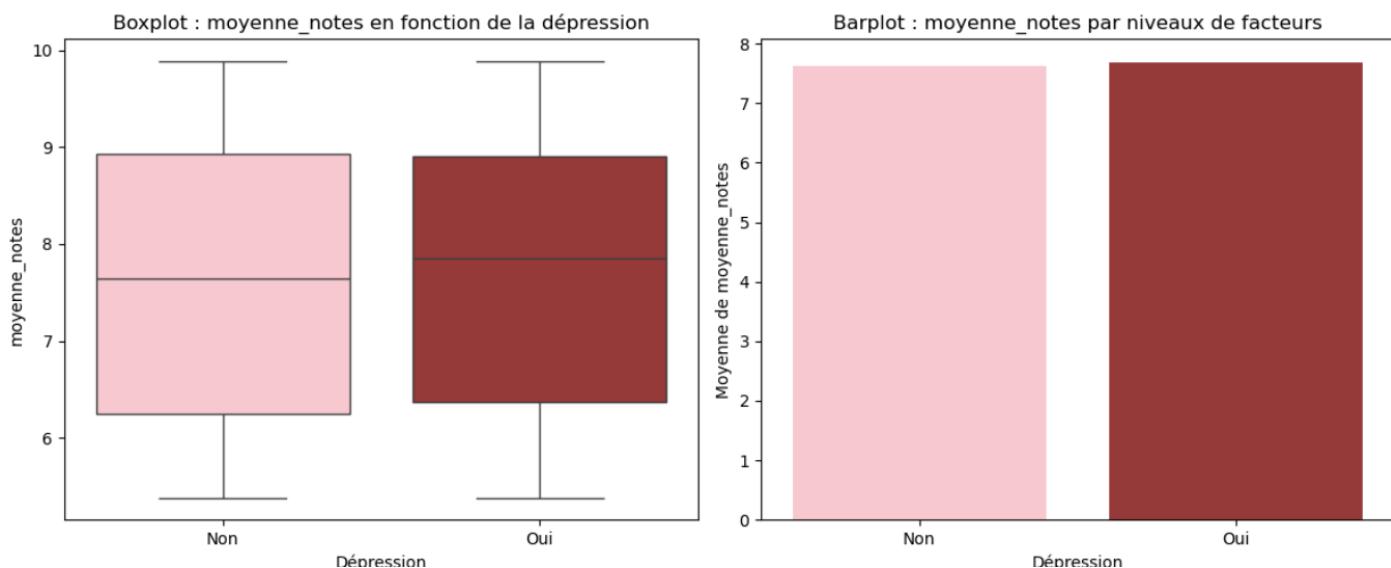
Pourquoi on ne peut pas utiliser certains tests ici :

1. **Manque de variation** : Les valeurs de pression liée au travail sont trop faibles et proches de zéro dans les deux groupes. Cela rend les tests statistiques inutiles car ils nécessitent une certaine dispersion ou variabilité dans les données pour détecter des différences.
2. **Appropriateness of tests** : Tests paramétriques comme le t-test ou l'ANOVA nécessitent des données avec une distribution normale et des variances significatives. Ici, les données ne répondent pas à ces conditions, donc ces tests ne seraient pas adaptés.

En conclusion, il est difficile d'utiliser un test statistique pour analyser la relation entre **pression liée au travail** et dépression étant donné l'absence de variation significative dans les données.

■ Analyse de la relation entre moyennes notes au travail et la dépression

La moyenne note ne suit pas une loi normale donc le test de kruskal est crucial

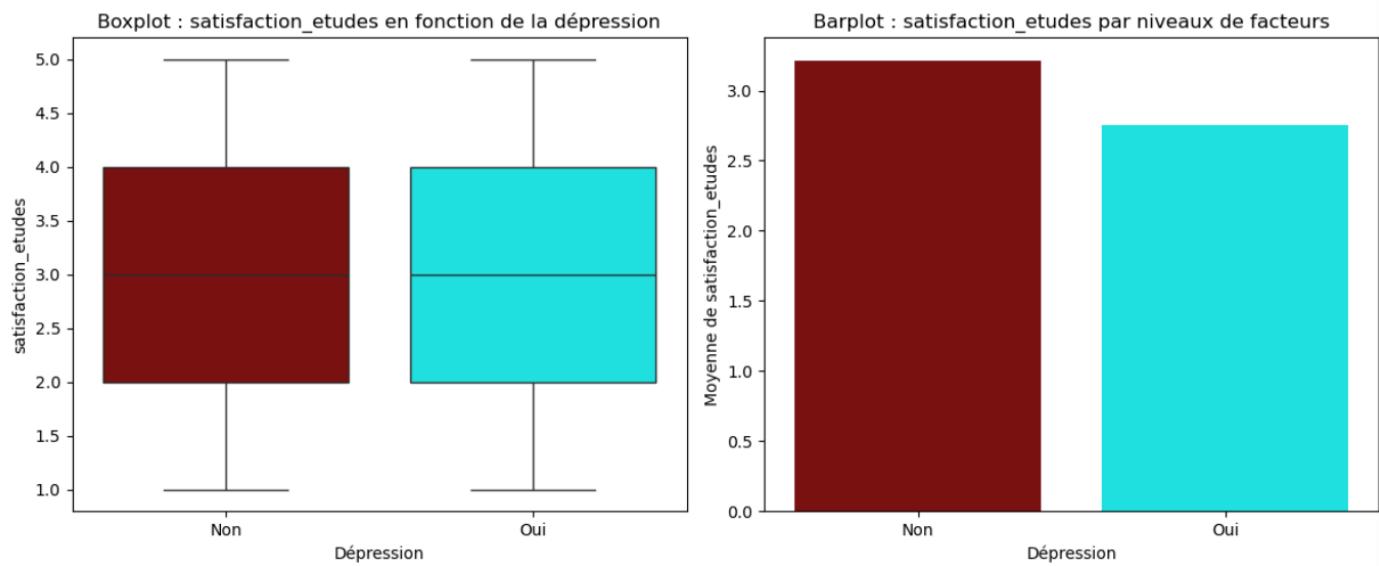


L'analyse de la relation entre les **moyennes des notes** et la dépression montre des résultats statistiquement significatifs. Le **boxplot** indique que les moyennes des notes des individus avec dépression (Oui) tendent à être légèrement supérieures à celles des individus sans dépression (Non), bien que la différence semble visuellement modérée. Le **barplot** confirme également une petite différence dans les moyennes.

Le test statistique de **Kruskal-Wallis**, avec une statistique de **13.170** et une **p-value de 0.0003**, indique une **relation significative** entre les moyennes des notes et la dépression. Bien que l'effet semble être modeste (différence visible mais faible), la significativité statistique confirme que cette relation ne peut être attribuée au hasard. Cela suggère que les moyennes des notes pourraient être influencées par l'état dépressif.

■ Analyse de la relation entre satisfaction études et la dépression

La satisfaction étude ne suit pas une loi normale donc le test de kruskal est crucial

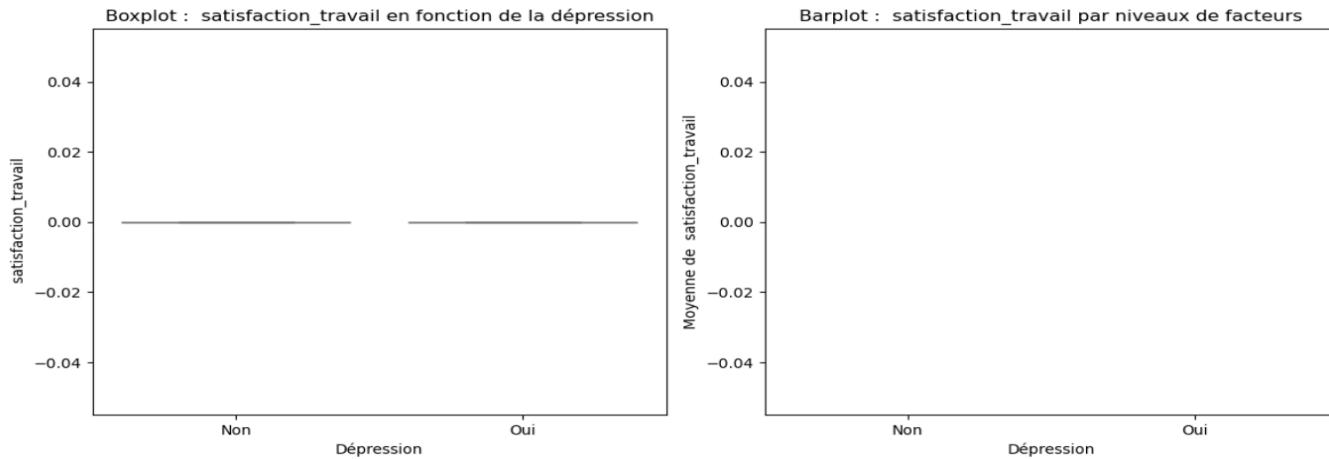


La relation entre la **satisfaction vis-à-vis des études** et la dépression est significative, comme le confirme le test de **Kruskal-Wallis** avec une statistique élevée (**789.519**) et une **p-value de 0.0000**.

Le **boxplot** montre que les individus dépressifs (Oui) tendent à avoir des niveaux de satisfaction vis-à-vis des études **plus faibles** comparés à ceux qui ne souffrent pas de dépression (Non). Le **barplot** renforce cette observation en affichant une moyenne de satisfaction inférieure pour les individus dépressifs.

Ces résultats suggèrent que la satisfaction vis-à-vis des études pourrait jouer un rôle clé dans la présence ou l'absence de dépression. Une exploration approfondie pourrait permettre de mieux comprendre la nature de cette association.

■ Analyse de la relation entre satisfaction travail et la dépression



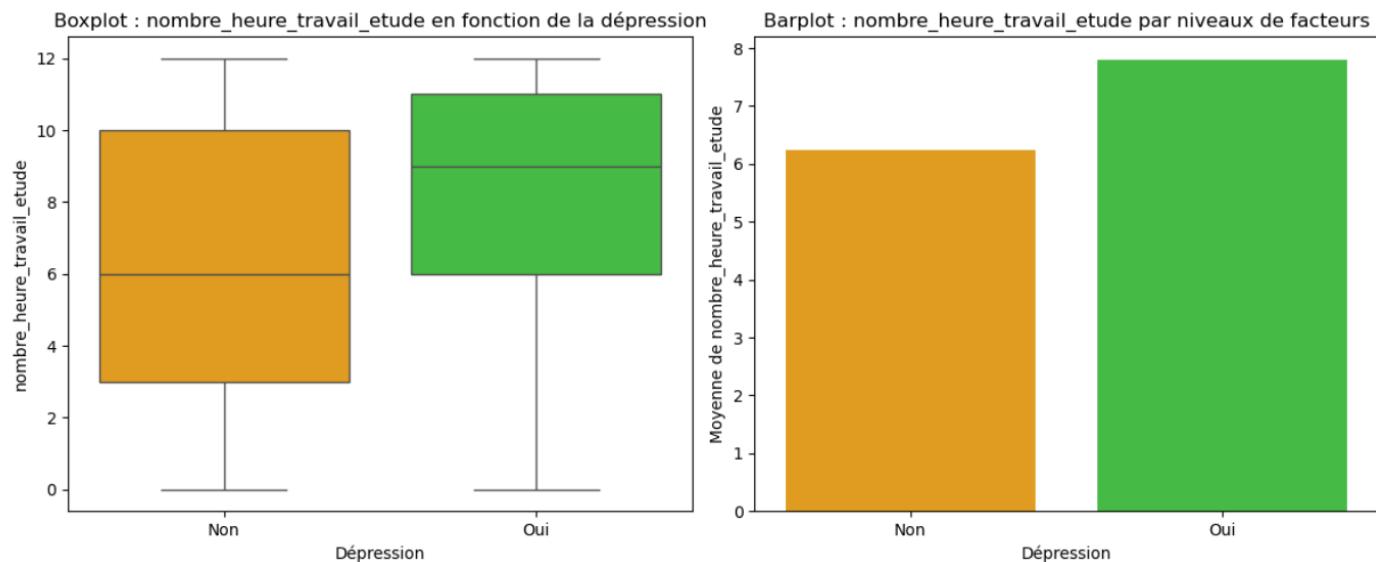
Les analyses montrent que la satisfaction au travail ne semble pas liée à la dépression.

Le boxplot indique une absence de variation significative dans la satisfaction au travail pour les deux groupes (Oui et Non), les médianes étant proches de zéro pour les deux. De plus, le barplot affiche une moyenne de satisfaction au travail également proche de zéro, indiquant que cette variable ne présente pas de différences significatives entre les niveaux de dépression.

Ainsi, la satisfaction au travail n'est pas un facteur explicatif pertinent pour la dépression dans cet ensemble de données.

- Analyse de la relation entre le nombre d'heure de travail ou étude et la dépression**

la variables nombre d'heure de travail ou études ne suit pas une loi normale donc le test de kruskal est crucial

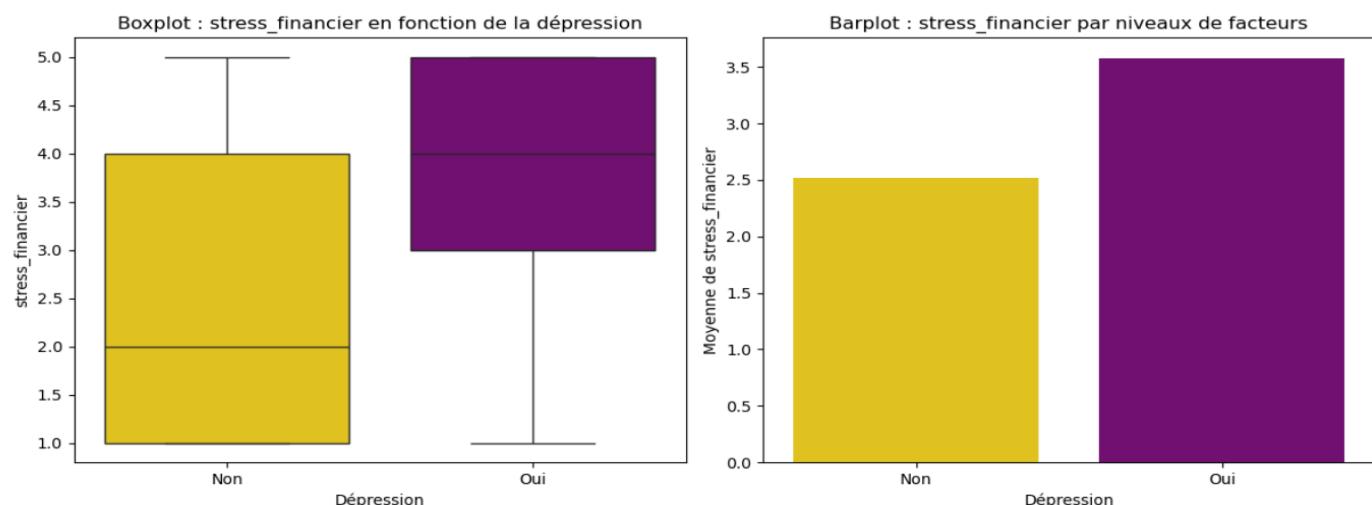


L'analyse montre une **relation significative** entre le **nombre d'heures de travail ou d'étude** et la dépression. Le **boxplot** indique que les individus souffrant de dépression tendent à consacrer **plus d'heures** au travail ou à l'étude, avec une médiane et une dispersion supérieures comparées à ceux sans dépression. Le **barplot** confirme que la moyenne d'heures est également plus élevée chez les individus dépressifs.

Le test de **Kruskal-Wallis** appuie cette observation avec une **statistique élevée (1124.132)** et une **p-value de 0.0000**, prouvant que la différence est significative et ne peut être attribuée au hasard. Ces résultats suggèrent que le nombre d'heures de travail ou d'étude pourrait jouer un rôle dans la prévalence de la dépression.

- Analyse de la relation entre le stress financier et la dépression**

Le stress financier ne suit pas une loi normale donc le test de kruskal est crucial



L'analyse révèle une **relation significative** entre le **stress financier** et la dépression. Le **boxplot** montre que les individus dépressifs (Oui) ont des niveaux de stress financier **plus élevés**, avec une médiane et une dispersion supérieures par rapport à ceux qui ne souffrent pas de dépression (Non). Le **barplot** confirme cette tendance, indiquant une moyenne de stress financier plus importante chez les individus dépressifs.

Le test de **Kruskal-Wallis**, avec une statistique élevée (**3672.102**) et une **p-value de 0.0000**, confirme que cette différence est statistiquement significative. Ces résultats suggèrent que le stress financier pourrait être un facteur clé associé à la dépression, nécessitant une attention particulière dans les analyses futures.

Résumé de l'analyse des variables quantitatives avec dépression

Dans cette première phase de l'analyse, nous avons exploré la relation entre la **dépression** (variable qualitative) et plusieurs **variables quantitatives** telles que l'âge, les moyennes des notes, le stress financier, et d'autres indicateurs. À l'aide de visualisations (boxplots et barplots), nous avons comparé les distributions et les moyennes des groupes dépressifs et non dépressifs. Ces observations ont été consolidées par le test statistique de **Kruskal-Wallis**, qui a permis d'évaluer si les différences observées étaient significatives. Les résultats ont révélé que certaines variables, comme le stress financier et le nombre d'heures de travail ou d'étude, sont fortement liées à la dépression, tandis que d'autres, comme la satisfaction au travail, n'ont pas montré de lien significatif.

Après cette analyse des variables quantitatives, nous allons maintenant examiner la relation entre la **dépression** et d'autres **variables qualitatives**. Cette étape permettra de déterminer si des associations significatives existent entre ces variables qualitatives et de mieux comprendre les facteurs pouvant être associés à l'état dépressif.

3. Pour deux variables qualitatives

La relation entre **deux variables qualitatives** consiste à examiner comment les catégories d'une variable sont associées aux catégories d'une autre. Cette analyse permet de comprendre si les deux variables sont indépendantes ou s'il existe une relation significative entre elles.

Calcul de l'effectif théorique de la variables dépression avec les autres variables qualitatives.

Variable	Effectifs théoriques minimaux	Condition de Cochran
sexe	5119.584773	<input checked="" type="checkbox"/> Respectée
profession	0.414474	<input type="checkbox"/> Non respectée
diplome_suivi	14.506595	<input checked="" type="checkbox"/> Respectée
pensees_suicidaire	4245.044304	<input checked="" type="checkbox"/> Respectée
antecedants_familiaux_maladie_mentale	5595.815578	<input checked="" type="checkbox"/> Respectée
ville	0.414474	<input type="checkbox"/> Non respectée
duree_sommeil	7.460535	<input checked="" type="checkbox"/> Respectée
habitudes_alimentaires	4.973690	<input type="checkbox"/> Non respectée

Variables respectant la condition de Cochran (test du Chi² fiable) :

- sexe, diplome_suivi, pensees_suicidaire, antecedants_familiaux_maladie_mentale, duree_sommeil

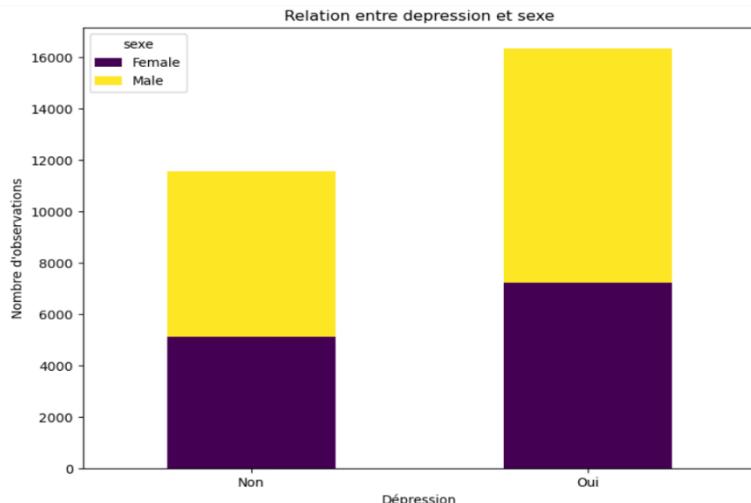
👉 Ces variables ont des effectifs théoriques suffisamment élevés : **on peut faire un test du Chi² fiable.**

Variables ne respectant pas la condition (test du Chi² non fiable) :

- profession, ville, habitudes alimentaires

👉 Ces variables ont des effectifs trop faibles : **le test du Chi² n'est pas valide**, les résultats pourraient être biaisés.

▪ Analyse de la relation entre le sexe et la dépression



Analyse du graphique

1. Barres représentant la dépression (Oui) et l'absence de dépression (Non) :

- ✓ Le graphique est divisé en deux catégories principales :
 - **Non** : Pas de dépression.
 - **Oui** : Dépression présente.
 - ✓ Chaque catégorie est subdivisée en deux segments :
 - **Homme** (jaune).
 - **Femme** (violet).

2. Observation des proportions :

- ✓ Il semble y avoir **plus de personnes avec dépression (Oui)**, indépendamment du sexe, que de personnes sans dépression (Non).
 - ✓ Pour les deux catégories, on observe une répartition similaire entre hommes et femmes.

Résultat statistique (Khi-Deux et p-value)

Le tableau associé indique la p-value pour la variable **sexe**, et le test de Khi-Deux :

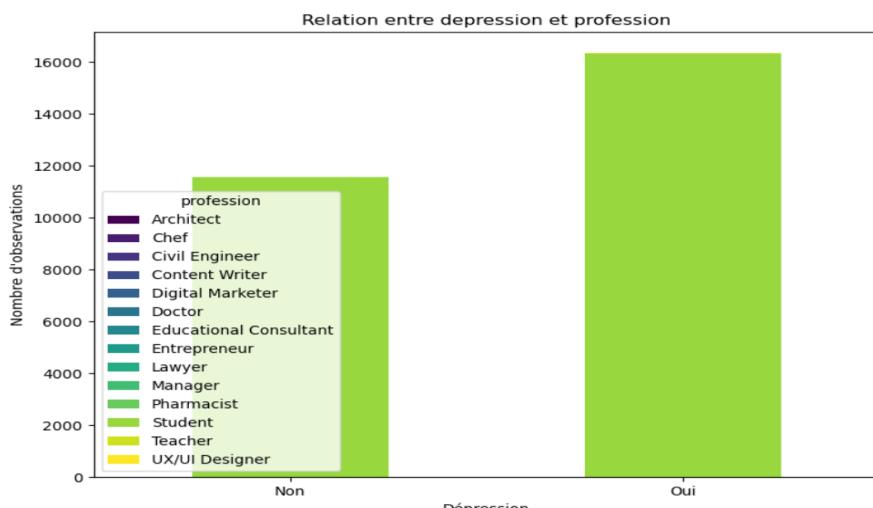
1. p-value = 0.770643 :

- ✓ Cette valeur est **supérieure à 0.05**, ce qui signifie que le résultat n'est pas statistiquement significatif au niveau classique de 5%.
 - ✓ **On ne rejette pas l'hypothèse nulle**, ce qui implique que le sexe et la dépression semblent indépendants (pas de lien significatif entre les deux).

2. Conclusion :

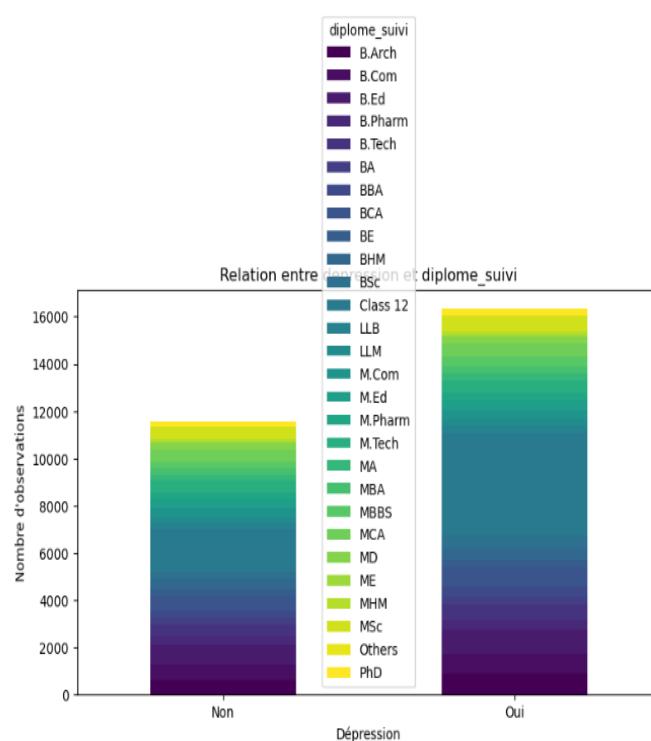
- ✓ Selon le test de Khi-Deux, il n'y a **pas d'association significative** entre le sexe et le fait d'avoir ou non une dépression dans les données analysées.

▪ Analyse de la relation entre la profession et la dépression



l'association entre profession et dépression.

■ Analyse de la relation entre la profession et la dépression



diplômes sont particulièrement liées à la dépression.

Ce résultat mérite une exploration approfondie pour identifier des tendances ou des profils spécifiques liés aux diplômes.

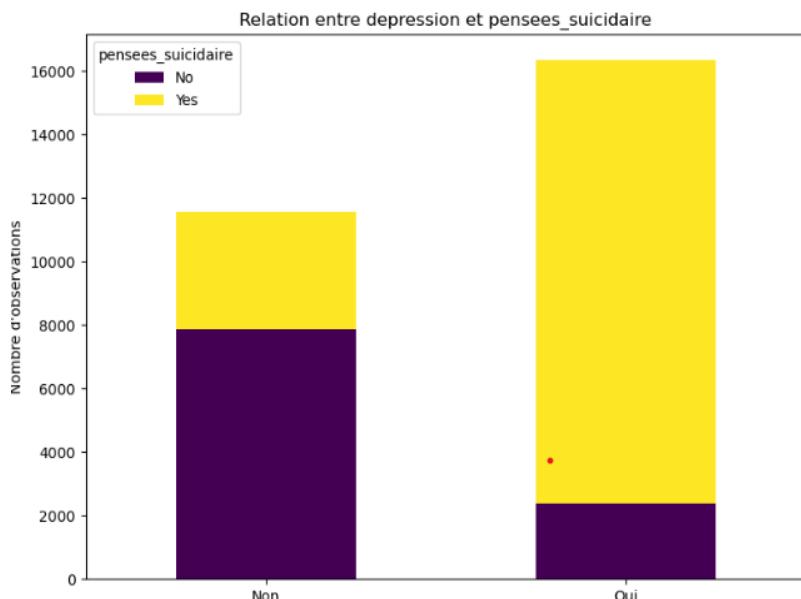
■ Analyse de la relation entre pensée suicidaire et la dépression

Le graphique montre que la dépression est majoritaire (16 000 individus) comparée à l'absence de dépression (8 000). Les professions incluent diverses catégories, mais il n'y a pas de test statistique adapté ici (Khi-Deux non valide selon Cochran, Fisher impossible avec plusieurs modalités). Une Analyse des Correspondances Multiples (ACM) ou l'indice de Cramér serait plus pertinent pour vérifier

Le test de **Khi-Deux** pour la variable **diplome_suivi** révèle une **p-value de 5.486092e-95**, soit un résultat extrêmement faible (inférieur à 0,05). Cela indique que l'association entre le diplôme suivi et la dépression est **statistiquement significative**. Autrement dit, le type de diplôme suivi influence fortement la présence ou l'absence de dépression parmi les individus étudiés.

Interprétation pratique :

- Il existe des **différences notables** dans les niveaux de dépression selon les diplômes suivis.
- Pour approfondir, une analyse plus détaillée (comme une Analyse des Correspondances Multiples ou une visualisation des proportions) serait utile pour comprendre quelles catégories de diplômes sont particulièrement liées à la dépression.



La **p-value de 0.000000** pour la variable **pensees_suicidaires** indique une **association extrêmement significative** entre les pensées suicidaires et la dépression. Cela signifie que les individus ayant des pensées suicidaires sont fortement liés à la présence de dépression.

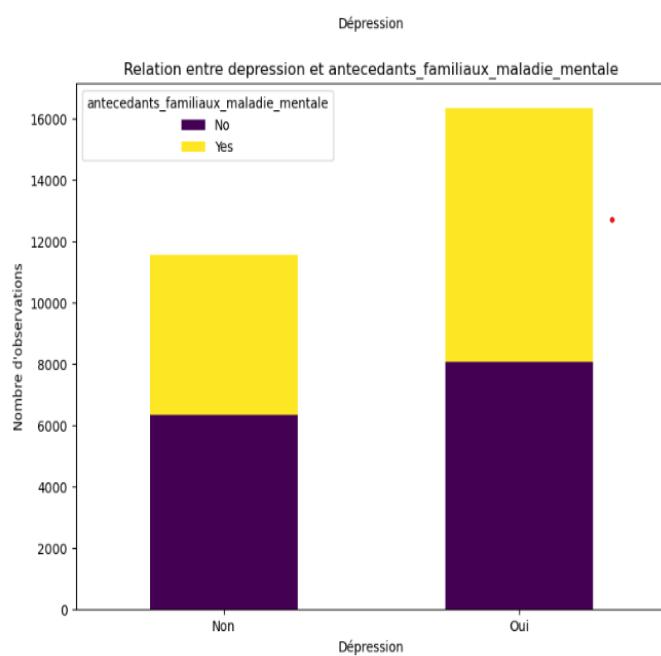
Interprétation du graphique :

Le graphique montre que :

- **Oui (dépression)** : La majorité des individus avec dépression ont également des pensées suicidaires.

- **Non (pas de dépression)** : Les pensées suicidaires sont beaucoup moins fréquentes chez les individus sans dépression.

- **Analyse de la relation entre antécédents familiaux de maladie mentale et la dépression**



La **p-value de 4.155503e-19** pour la variable **antécédents familiaux de maladie mentale** est extrêmement faible (bien en dessous de 0,05). Cela signifie qu'il existe une **association significative** entre les antécédents familiaux de maladie mentale et la dépression.

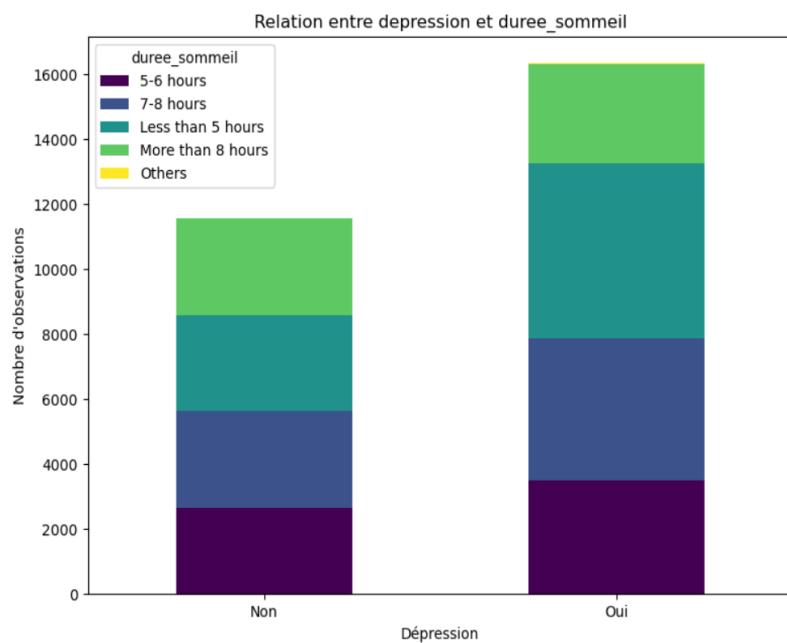
Interprétation pratique :

- Les individus ayant des antécédents familiaux de maladies mentales (Oui) sont fortement associés à la présence de dépression.
- Le graphique montre que :
 - La catégorie **Oui (dépression)** a une proportion élevée d'individus avec des antécédents familiaux.
 - La catégorie **Non (pas de dépression)** est moins liée à ces antécédents.

Conclusion :

Ces résultats soulignent l'importance de considérer l'histoire familiale dans l'évaluation des risques de dépression. Une analyse plus approfondie pourrait explorer les modalités spécifiques des maladies mentales familiales influençant la dépression.

■ Analyse de la relation entre diplôme suivi et la dépression



La **p-value** de **9.240458e-59** pour la variable **durée de sommeil** est extrêmement faible (bien inférieure à 0,05), ce qui signifie que la relation entre la durée de sommeil et la dépression est **hautement significative**.

Interprétation :

- Le graphique montre une répartition des catégories de durée de sommeil (moins de 5 heures, 5-6 heures, 7-8 heures, plus de 8 heures, et autres) parmi les individus dépressifs (Oui) et non dépressifs (Non).
- Les individus avec une **durée de sommeil anormale** (moins de 5 heures ou plus de 8 heures) semblent plus représentés dans la catégorie "Oui" (dépression).

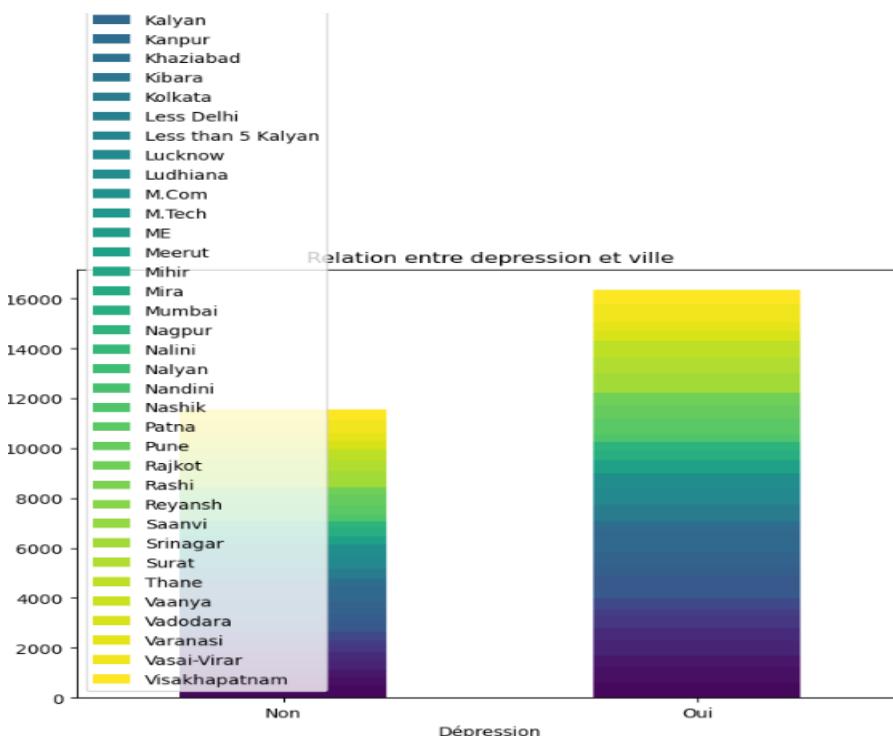
catégorie "Oui" (dépression).

- Une durée de sommeil "normale" (7-8 heures) semble être plus fréquente chez les individus "Non" (pas de dépression).

Conclusion :

La durée de sommeil est un facteur clé associé à la dépression. Les personnes ayant une durée de sommeil trop courte ou trop longue sont plus susceptibles d'être dépressives, ce qui souligne l'importance d'un sommeil équilibré pour la santé mentale.

■ Analyse de la relation entre ville et la dépression



La relation entre la variable "**ville**" et la "**dépression**" ne peut pas être testée statistiquement avec le **test de Fisher**, car ce dernier n'est applicable qu'à des tables 2x2 (deux modalités pour chaque variable). De plus, la règle de **Cochran** n'est pas respectée : cela signifie que certaines fréquences dans le tableau croisé sont trop faibles pour que le **test du Chi-Deux** soit valide.

Interprétation générale :

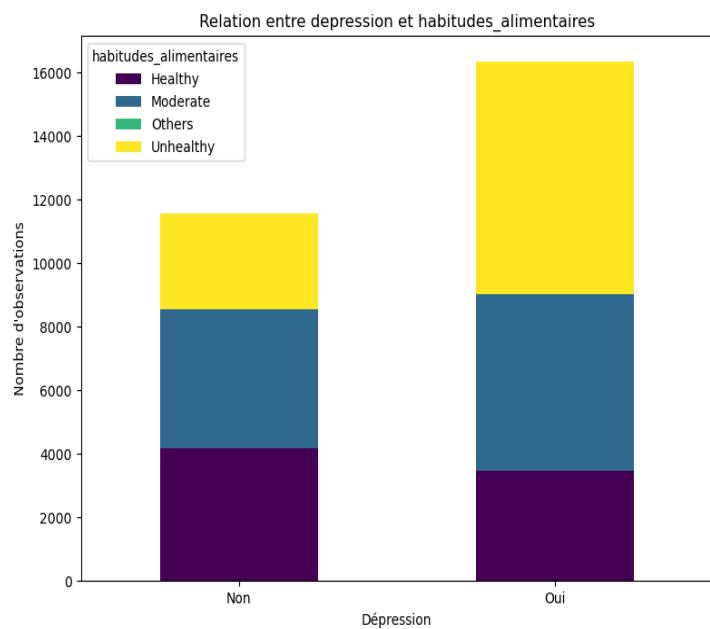
Cependant, l'analyse visuelle du graphique montre des proportions de dépression ("Oui") et de non-dépression ("Non") réparties en fonction des villes. Cela indique :

- Une possible variation géographique entre les villes en termes de dépression.
- Certaines villes pourraient avoir des populations davantage affectées par la dépression que d'autres, mais cette observation reste qualitative et doit être explorée avec des méthodes adaptées.

Approches possibles pour approfondir l'analyse :

1. **Analyse des Correspondances Multiples (ACM)** : Cette méthode est idéale pour explorer des relations entre variables qualitatives comportant plusieurs modalités.
2. **Regroupement (clustering)** : Si les villes peuvent être regroupées en clusters (par caractéristiques similaires), cela pourrait révéler des tendances géographiques intéressantes.
3. **Test exact multinomial** : Une alternative statistique pouvant être utilisée lorsque Fisher ou Chi-Deux ne sont pas applicables.

Analyse de la relation entre ville et la dépression



La variable **habitudes alimentaires** ne peut pas être testée avec le **test exact de Fisher**, car ce dernier est limité aux tables 2x2. De plus, le non-respect de la règle de **Cochran** indique que certaines fréquences sont trop faibles pour appliquer le test de Chi-Deux. Cela signifie qu'aucun de ces tests statistiques classiques n'est approprié pour analyser la relation entre les **habitudes alimentaires** et la **dépression**.

Interprétation générale :

Le graphique montre une tendance où les **habitudes alimentaires malsaines** sont plus fréquentes chez les individus dépressifs (**Oui**) que chez ceux qui ne sont pas dépressifs (**Non**).

Cela suggère une possible corrélation entre des habitudes alimentaires déséquilibrées et la dépression, mais cette observation reste qualitative sans validation statistique.

Approches alternatives pour cette analyse :

1. **Analyse des Correspondances Multiples (ACM)** :
 - Permet d'étudier les associations entre plusieurs modalités qualitatives.
 - Visualise la proximité entre les catégories "saines", "malsaines", etc., et la variable de dépression.
2. **Analyse Factorielle des Correspondances (AFC)** :
 - Explore les relations dans un tableau croisé, idéal pour des données qualitatives.
3. **Clustering** :

- Regroupe les individus selon leurs habitudes alimentaires et leur état de dépression pour identifier des profils.

Après avoir exploré les relations **bivariées**, il est temps de passer à l'étape suivante : une analyse **multivariée**. Cette phase permettra de mieux comprendre les interactions complexes entre plusieurs variables, d'identifier des structures sous-jacentes ou des regroupements significatifs, et d'obtenir une vue d'ensemble plus riche et approfondie des données.

3 ème PARTIE : ANALYSE DE DONNEES MULTIVARIEES

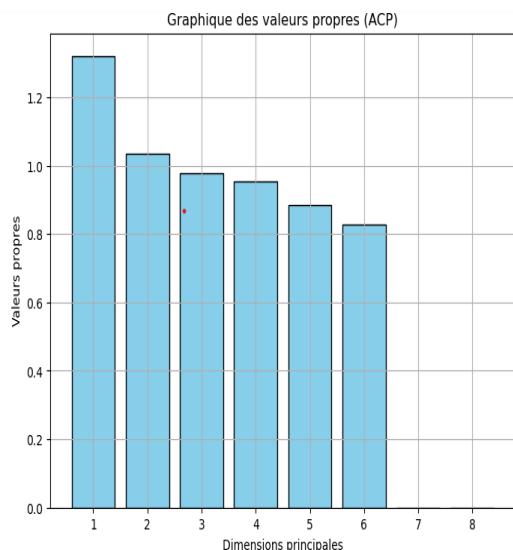
L'**analyse de données multivariées** étudie les relations complexes entre plusieurs variables simultanément. Elle vise à réduire les dimensions, identifier des regroupements (clustering) ou explorer des interactions. Les méthodes courantes incluent l'**ACP** (pour les variables quantitatives), l'**ACM/AFC** (pour les variables qualitatives) et les techniques de **clustering** pour regrouper des individus ou modalités similaires. Cela offre une vision globale des données pour mieux comprendre leur structure

1. Méthodes factorielles

Les **méthodes factorielles** sont des techniques d'analyse multivariée utilisées pour identifier des structures ou des relations cachées dans les données. Elles se concentrent sur la réduction de dimension tout en conservant un maximum d'information. Voici un aperçu :

ACP (Analyse en Composantes Principales) : Utilisée pour des variables quantitatives, elle réduit les données en un petit nombre de dimensions principales tout en expliquant la majorité de la variance

- Graphique des valeurs propre



Interprétation globale

Ce graphique représente les valeurs propres obtenues lors d'une Analyse en Composantes Principales (ACP). Voici une interprétation concise :

1. Première dimension (Dimension 1) : Elle a la plus grande valeur propre, ce qui signifie qu'elle explique le plus de variance dans les données. Elle est donc la dimension la plus importante.
2. Réduction progressive : Les valeurs propres diminuent au fur et à mesure que l'on avance vers les dimensions suivantes, indiquant que chaque dimension supplémentaire explique une part de variance de plus en plus faible.
3. Saturation : À partir d'une certaine dimension (probablement la 3ème ou la 4ème, selon l'échelle), les valeurs propres sont très faibles, ce qui montre qu'elles ajoutent peu d'information significative.

Conclusion

Ce graphique suggère qu'on peut se concentrer sur les **trois premières composantes** pour les analyses ultérieures, car elles captent la majorité de l'information, réduisant ainsi la complexité des données sans perte significative

➤ Interprétation des inerties

```
Inerties cumulées (% d'information captée) :
Dimension 1: 22.02%
Dimension 2: 39.29%
Dimension 3: 55.59%
Dimension 4: 71.48%
Dimension 5: 86.20%
Dimension 6: 100.00%
Dimension 7: 100.00%
Dimension 8: 100.00%
```

Ce tableau montre les **inerties cumulées** exprimées en pourcentage, révélant combien d'information chaque dimension factorielle capture dans l'analyse. Voici l'interprétation :

- Dimension 1** : Elle capture **22.02%** de l'information totale, la plus significative parmi toutes les dimensions.
- Ajout progressif des dimensions** : La **Dimension 2** porte le cumul à **39.29%**, la **Dimension 3** à **55.59%**, et ainsi de suite. Chaque nouvelle dimension ajoute de l'information, mais de façon décroissante.
- Saturation** : À partir de la **Dimension 6** (100% d'information captée), aucune nouvelle information n'est ajoutée. Cela indique que les six premières dimensions suffisent pour expliquer entièrement les données.

Conclusion :

Pour simplifier l'analyse, il est pertinent de se concentrer sur les **premières dimensions** (1 à 3), puisqu'elles capturent la majeure partie de l'information tout en réduisant la complexité des données.

➤ Corrélations entre les variables et les dimensions

Variable	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8
age	-0.344	0.350	0.696	0.044	0.522	0.014	~0	~0
pression_academique	0.571	-0.047	0.204	-0.176	0.129	0.763	~0	~0
pression_liée_au_travail	~0	~0	~0	~0	~0	~0	0.147	0.989
moyenne_notes	0.016	0.759	-0.503	0.322	0.139	0.220	~0	~0
satisfaction_etudes	-0.362	-0.539	-0.264	0.458	0.427	0.341	~0	~0
satisfaction_travail	~0	~0	~0	~0	~0	~0	0.989	-0.147
nb_heure_travail_etude	0.372	-0.031	0.348	0.807	-0.261	-0.143	~0	~0
stress_financier	0.535	-0.090	-0.176	-0.045	0.664	-0.482	~0	~0

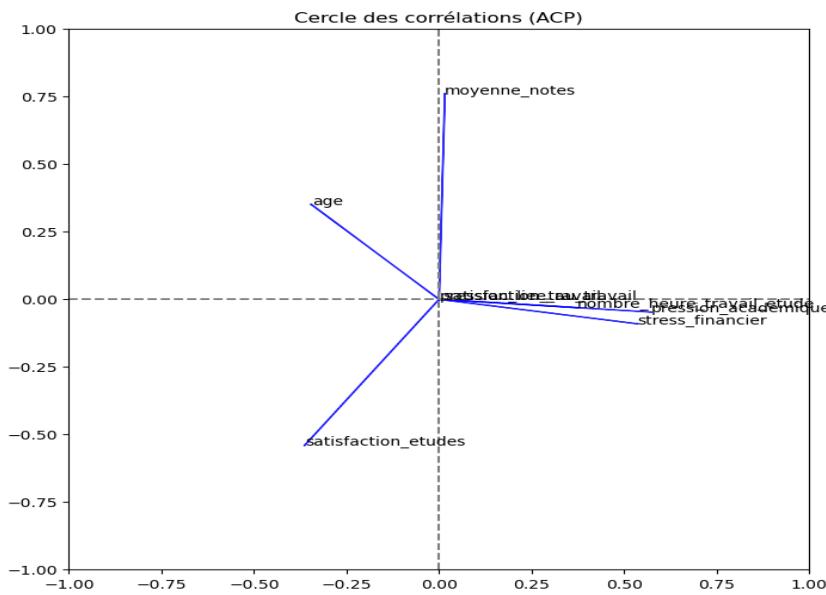
Les données montrent les contributions de chaque variable à différentes **dimensions factorielles**, mettant en lumière leur influence dans une analyse multivariée (comme une ACP ou ACM). Voici les observations principales :

1. Variables dominantes :

- Age est fortement corrélé à la **Dimension 3** (0.696), ce qui suggère que cette dimension est en grande partie liée à l'âge.
- Pression académique a une contribution significative dans la **Dimension 1** (0.571) et dans la **Dimension 6** (0.763).

- Stress financier se distingue dans la **Dimension 1** (0.535) et la **Dimension 5** (0.664).
2. Spécificité des variables :
 - Pression liée au travail et satisfaction au travail semblent liées aux dernières dimensions (**Dim 7 et Dim 8**), avec des valeurs élevées proches de **0.989** et **-0.147**.
 3. Impacts combinés :
 - Les variables comme **nombre d'heures de travail ou d'étude** et **satisfaction aux études** montrent des contributions variées sur plusieurs dimensions, reflétant leurs relations complexes dans les données.

➤ Cercle de corrélation



Ce graphique montre un **cercle des corrélations** provenant d'une **Analyse en Composantes Principales (ACP)**. Voici l'interprétation :

1. Relation entre les variables

- Les variables représentées par des vecteurs proches les unes des autres (comme **stress financier** et **pression académique**) sont positivement corrélées. Cela signifie qu'elles ont tendance à varier ensemble dans les données.
- Des vecteurs opposés, comme **satisfaction au travail** et **stress financier**, suggèrent une corrélation

négative. Ces deux variables évoluent dans des directions opposées.

2. Contribution aux dimensions

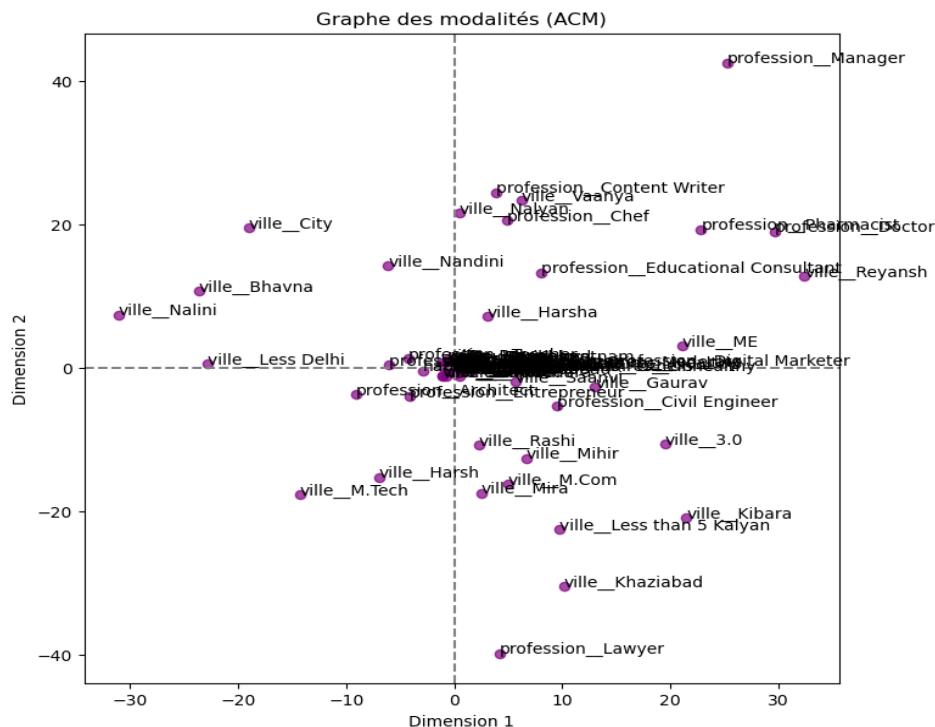
- **Dim1** et **Dim2** sont les axes principaux issus de l'ACP. Les variables dont les vecteurs sont proches de ces axes contribuent fortement à expliquer la variance captée par ces dimensions. Par exemple :
 - **Satisfaction aux études** contribue à **Dim1**.
 - **Moyenne des notes** est fortement associée à **Dim2**.

3. Interprétation pratique

Ce cercle illustre la manière dont les variables interagissent dans l'espace factoriel.

- **ACM (Analyse des Correspondances Multiples)** : Conçue pour les variables qualitatives, elle explore les relations entre catégories et visualise leurs associations.

➤ Graphe



groupées autour d'une zone partagent probablement des attributs ou comportements similaires avec les villes correspondantes.

2. Contributions des dimensions :

- **Dimension 1 et Dimension 2** capturent les principales variations entre les modalités. Les modalités éloignées des axes sont celles qui contribuent davantage à expliquer ces variations.

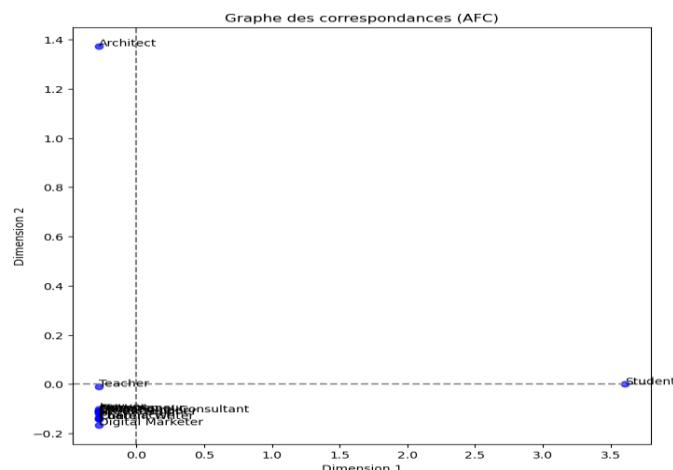
3. Clusters visibles :

- Certaines modalités forment des regroupements naturels, ce qui peut indiquer des profils ou associations spécifiques entre profession, ville, et autres variables qualitatives.

Conclusion :

Ce graphique est utile pour identifier des corrélations ou regroupements significatifs dans les données. Il met en évidence les associations fortes ou faibles entre les modalités étudiées, facilitant des

AFC (Analyse Factorielle des Correspondances) : Une extension des tableaux croisés, utile pour analyser les liens entre deux variables qualitatives.



Ce graphique issu de l'**Analyse Factorielle des Correspondances (AFC)** illustre la position des professions dans un espace factoriel défini par deux dimensions principales (**Dimension 1 et Dimension 2**). Voici une interprétation concise :

1. Positionnement des professions

- **Architecte** : Placé haut sur la **Dimension 2** (à environ 0.0, 1.4), ce qui indique une distinction significative par rapport aux autres professions dans cette dimension.
- **Enseignant** : Situé au centre (0.0, 0.0), ce qui suggère une position équilibrée, sans forte contribution aux deux dimensions principales.
- **Étudiant** : Loin sur la **Dimension 1** (3.5, 0.0), ce qui montre que cette profession est fortement corrélée avec cette dimension.
- **Manager, Digital Marketer, et Ingénieur** : Ces professions sont regroupées autour de (0.0, -0.2), ce qui indique qu'elles partagent des similarités dans leurs contributions aux deux dimensions.

2. Interprétation des dimensions

- **Dimension 1** semble capturer des éléments liés à l'activité ou aux rôles opérationnels, avec l'étudiant fortement corrélé.
- **Dimension 2** pourrait refléter des aspects liés aux qualifications ou au niveau de responsabilité, mettant en avant l'architecte.

3. Regroupement et dispersion

Les professions regroupées montrent des similarités ou une connexion dans les critères sous-jacents, tandis que celles éloignées, comme l'architecte et l'étudiant, révèlent des différences marquées dans leurs relations avec les dimensions étudiées.

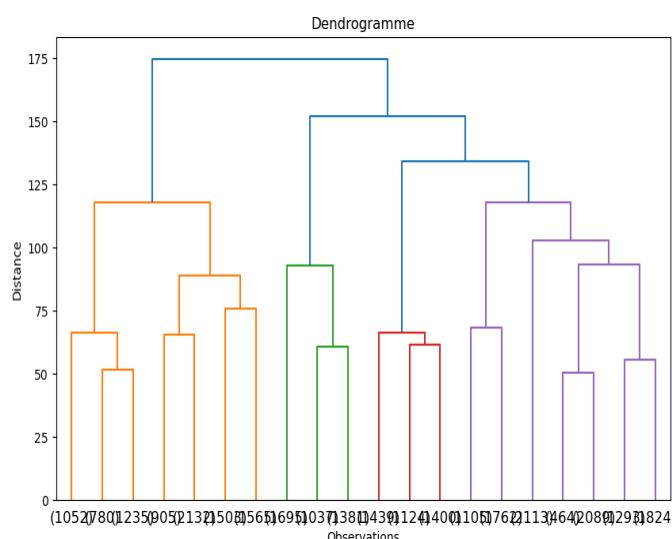
Conclusion

Ce graphique permet d'explorer les relations et divergences entre les professions. Il fournit une base pour identifier les clusters ou analyser les dimensions qui influencent les attributs spécifiques des professions.

Clustering

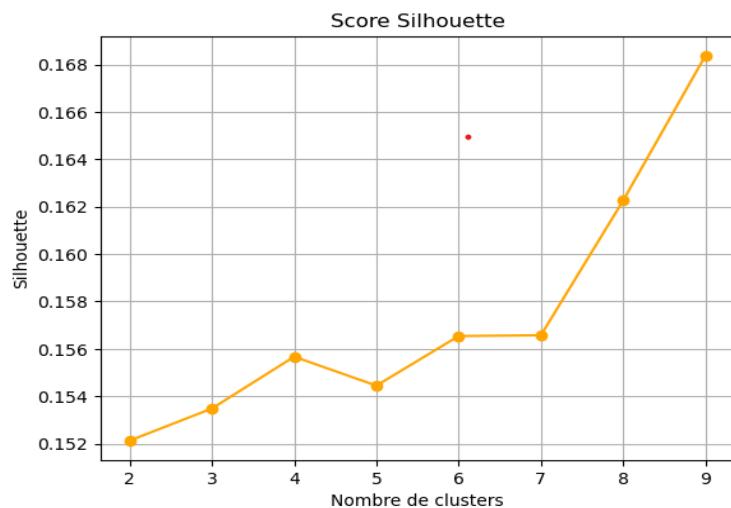
Le **clustering** regroupe les données similaires en **clusters** pour révéler des patterns cachés. Par exemple, on peut classer des professions en groupes selon leur stress ou satisfaction. Des algorithmes comme **K-Means** ou **DBSCAN** sont courants. Cela aide à analyser et à comprendre des profils ou segments spécifiques dans les données

- Dendrogramme



Le dendrogramme visualise la hiérarchie de regroupement des données. Les branches montrent comment les observations ou groupes se rapprochent progressivement selon leur similarité. La hauteur des fusions représente la dissimilarité : plus elle est basse, plus les observations sont similaires. En définissant un seuil sur l'axe des distances, on peut découper les données en clusters distincts, permettant d'identifier des groupes homogènes pour une analyse ciblée. Ces clusters sont utiles pour comprendre des patterns cachés ou explorer les relations entre les données.

➤ Détermination du nombre optimal de clusters



Ce graphique, intitulé **Score Silhouette**, évalue la qualité des regroupements pour différents nombres de clusters. Voici l'interprétation :

1. Axe des clusters et silhouette

- L'axe horizontal représente le **nombre de clusters** (de 2 à 9).
- L'axe vertical montre le **score silhouette**, qui mesure la cohésion et la séparation des clusters. Un score élevé indique des clusters bien définis.

2. Tendance générale

- On observe une augmentation globale du score silhouette lorsque le nombre de clusters augmente, suggérant une meilleure segmentation à mesure que les clusters se multiplient.

3. Point remarquable

- Un pic est clairement visible au niveau de **9 clusters**, où le score atteint sa valeur maximale. Cela pourrait indiquer que **9 clusters** est un bon choix pour segmenter ces données.

Conclusion : Ce graphique aide à choisir le nombre optimal de clusters pour un algorithme de classification. Le pic à 9 clusters suggère une solution bien structurée, tandis que les autres valeurs offrent une segmentation moins cohérente.

➤ Caractérisations des clusters

Cluster	Pression Académique	Pression liée au Travail	Moyenne Notes	Satisfaction Études	Satisfaction Travail	Nombre d'Heures Travail/Étude	Stress Financier
0	4.153265	0.0	6.884993	1.849828	0.0	9.044330	1.856357
1	3.045018	0.0	8.803313	4.227491	0.0	7.510504	4.312725
2	4.118148	0.0	8.785522	1.744279	0.0	9.436136	4.111495
3	2.583275	0.0	9.007808	2.829617	0.0	2.500348	2.048780
4	3.981777	0.0	7.062107	2.015566	0.0	2.877752	4.223994
5	1.422441	0.0	7.436916	1.916286	0.0	8.808653	3.357017
6	2.276868	0.0	6.297751	3.909964	0.0	3.095730	2.201423
7	3.837776	0.0	6.201697	3.762396	0.0	9.419031	4.134422
8	2.522755	0.0	8.141688	4.035978	0.0	9.682657	1.602706

Ce tableau résume les données des clusters en fonction de variables comme la pression académique, la satisfaction et le stress. Voici une analyse concise :

1. Cluster 0

- **Pression Académique** : Moyenne (4.15), ce qui indique une charge de travail modérée.
- **Moyenne Notes** : Assez basse (6.88), suggérant des difficultés à performer académiquement.

- **Stress Financier** : Faible (1.85), indiquant peu de pression financière.
- **Nombre d'Heures de Travail/Étude** : Élevé (9.04), ce qui montre un engagement significatif dans les études.
- **Satisfaction Études** : Faible (1.84), les étudiants semblent insatisfaits malgré leurs efforts.

2. Cluster 1

- **Pression Académique** : Relativement faible (3.04).
- **Moyenne Notes** : Haute (8.80), suggérant une bonne performance académique.
- **Stress Financier** : Élevé (4.31), ce groupe ressent une pression financière notable.
- **Satisfaction Études** : Haute (4.22), bien que sous tension financière, ils sont satisfaits de leurs études.
- **Nombre d'Heures Travail/Étude** : Moyenne (7.51).

3. Cluster 2

- **Pression Académique** : Moyenne (4.11).
- **Moyenne Notes** : Haute (8.78), similaire au Cluster 1.
- **Stress Financier** : Élevé (4.11), ce qui pourrait affecter leur équilibre.
- **Satisfaction Études** : Faible (1.74), malgré de bons résultats, ils montrent une insatisfaction notable.
- **Nombre d'Heures Travail/Étude** : Élevé (9.43).

4. Cluster 3

- **Pression Académique** : Faible (2.58).
- **Moyenne Notes** : Très haute (9.00), un excellent niveau de performance.
- **Stress Financier** : Faible (2.04).
- **Satisfaction Études** : Moyenne (2.82).
- **Nombre d'Heures Travail/Étude** : Très faible (2.50), suggérant une méthode de travail efficace.

5. Cluster 4

- **Pression Académique** : Moyenne (3.98).
- **Moyenne Notes** : Moyenne (7.06).
- **Stress Financier** : Élevé (4.22), ce groupe semble lutter avec des contraintes financières.
- **Satisfaction Études** : Moyenne (2.01).
- **Nombre d'Heures Travail/Étude** : Faible (2.87).

7. Cluster 5

- **Pression Académique** : Très faible (1.42).
- **Moyenne Notes** : Moyenne (7.43).
- **Satisfaction Études** : Faible (1.91).
- **Stress Financier** : Modéré (3.35).
- **Heures de travail/étude** : Élevées (8.80).
- **Interprétation** : Profil d'individus peu pressurisés mais investissant beaucoup de temps dans leurs études, avec une satisfaction limitée.

7. Cluster 6

- **Pression Académique** : Faible (2.27).
- **Moyenne Notes** : Basse (6.29).
- **Stress Financier** : Faible (2.20).
- **Satisfaction Études** : Très haute (3.90), ce groupe est satisfait malgré des résultats académiques modestes.
- **Nombre d'Heures Travail/Étude** : Faible (3.09).

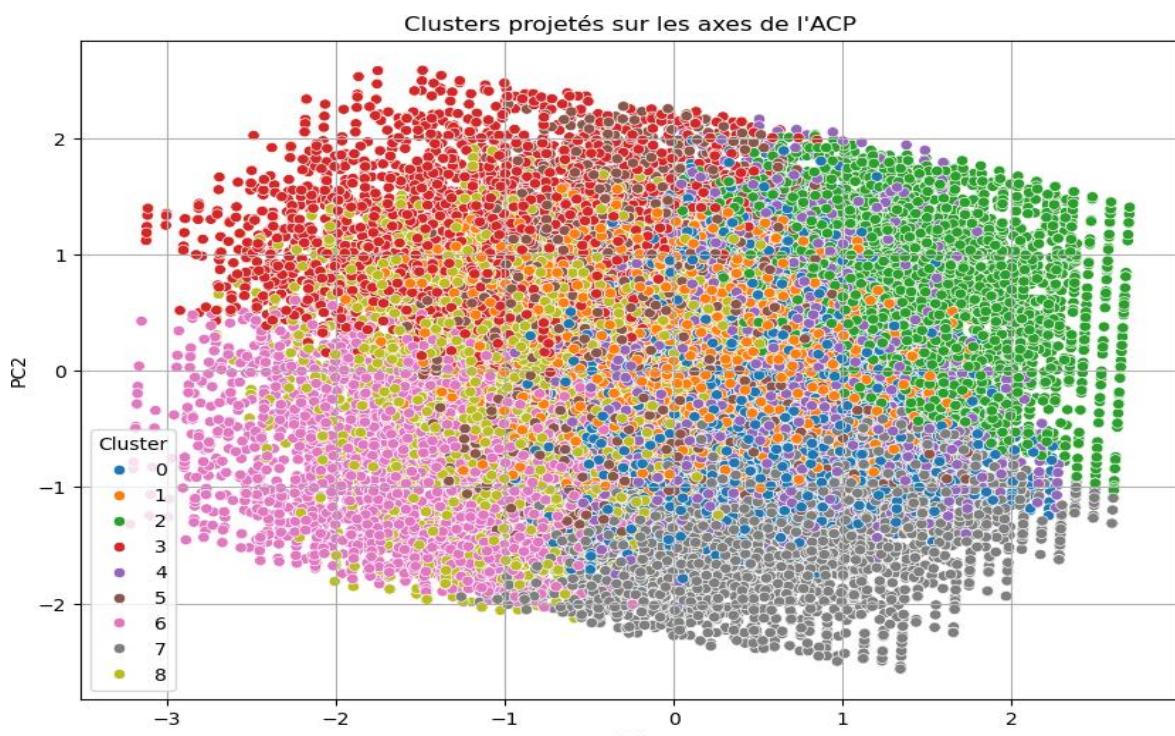
8. Cluster 7

- **Pression Académique** : Moyenne (3.83).
- **Moyenne Notes** : Basse (6.20).

- **Stress Financier** : Élevé (4.13).
- **Satisfaction Études** : Haute (3.76).
- **Nombre d'Heures Travail/Étude** : Élevé (9.41), ils travaillent beaucoup mais leurs résultats restent faibles.

9. Cluster 8

- **Pression Académique** : Faible (2.52).
- **Moyenne Notes** : Moyenne (8.14).
- **Stress Financier** : Très faible (1.60).
- **Satisfaction Études** : Très haute (4.03).
- **Nombre d'Heures Travail/Étude** : Très élevé (9.68), ce groupe est très engagé et satisfait de ses études.



4 -ème parties : réponses aux questions

1. Quel est l'intervalle de confiance pour la proportion d'étudiants ayant déjà eu des pensées suicidaires ?

- L'intervalle de confiance pour la proportion d'étudiants ayant déjà eu des pensées suicidaires se situe entre **0,6272 et 0,6385**. Cela signifie que, avec un certain niveau de confiance (souvent 95 %), on estime que la proportion réelle d'étudiants ayant exprimé de telles pensées se trouve dans cet intervalle. Cet outil statistique permet d'évaluer la précision de l'estimation et d'apporter une meilleure compréhension de la prévalence de cette problématique dans la population étudiante étudiée.

2. Moyenne et Médiane :

❖ Estimez la moyenne et la médiane des heures de travail ou d'études pour les étudiants souffrant de dépression.

La moyenne des heures de travail ou d'études pour les étudiants souffrant de dépression est estimée à **7.81 heures**, tandis que la médiane est légèrement plus élevée, à **9.00 heures**. Ces chiffres montrent que, malgré la dépression, une part significative des étudiants consacre une quantité importante de temps au travail.

ou aux études. La médiane, supérieure à la moyenne, suggère une asymétrie dans la distribution des heures, avec certains étudiants travaillant ou étudiant nettement plus que la majorité.

- ❖ Évaluez la moyenne et la médiane du stress financier pour les étudiants avec et sans dépression.

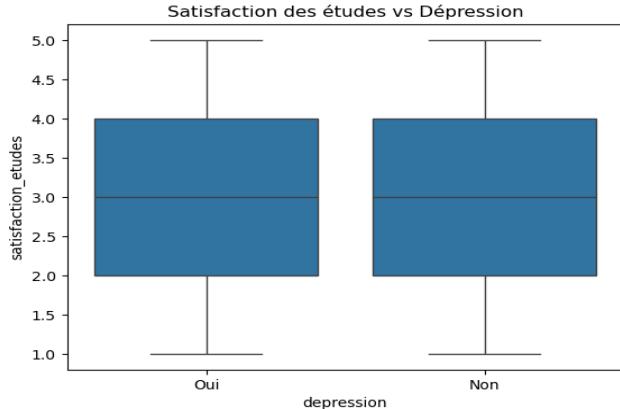
Pour les étudiants souffrant de dépression, la **moyenne** du stress financier est estimée à **3.58**, tandis que la **médiane** est légèrement plus élevée, à **4.0**. Cela suggère que, globalement, cette population ressent un niveau de stress financier significatif, avec certains cas encore plus marqués.

En comparaison, pour les étudiants **sans dépression**, la **moyenne** du stress financier est de **2.52**, et la **médiane** est de **2.0**, indiquant un stress financier relativement plus faible pour cette population.

Cette différence notable entre les deux groupes reflète une corrélation possible entre la dépression et une plus grande pression financière, nécessitant potentiellement des interventions ciblées pour soulager ce stress parmi les étudiants affectés.

3. Différence de moyennes :

- ❖ La satisfaction des études diffère-t-elle significativement entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas ?



La satisfaction des études diffère significativement entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas. Les résultats montrent que :

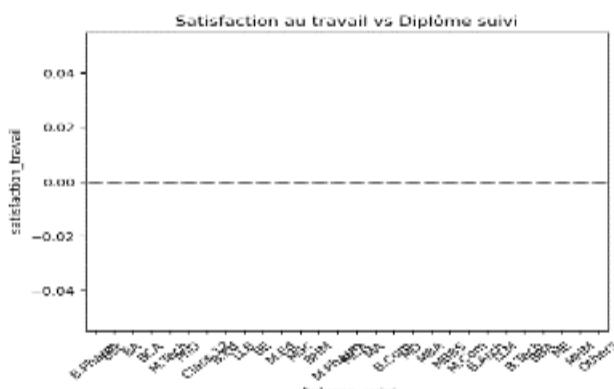
- **Étudiants sans dépression** : Moyenne de satisfaction élevée à **3.22** avec un écart-type de **1.33**.
- **Étudiants avec dépression** : Moyenne de satisfaction plus basse à **2.75**, avec un écart-type de **1.35**.

Les tests de normalité (pour les deux groupes) révèlent des distributions non normales ($p = 0.0000$). En conséquence, le test de **Mann-Whitney** a été utilisé, confirmant une différence statistiquement significative

($p\text{-value} = 0.0000$) dans la satisfaction entre les deux groupes.

Conclusion : La dépression est associée à une satisfaction des études réduite, ce qui peut refléter un impact émotionnel et psychologique important chez ces étudiants. Ces résultats mettent en lumière l'importance de l'accompagnement psychologique dans l'environnement académique.

- ❖ Les niveaux de satisfaction au travail diffèrent-ils significativement selon le diplôme suivi ?



Les niveaux de satisfaction au travail ne diffèrent pas significativement selon le diplôme suivi. Le message d'erreur **ValueError: All numbers are identical in kru** indique que toutes les valeurs de satisfaction au travail pour les différents diplômes sont identiques (celles-ci semblent être centrées autour de 0), ce qui empêche tout test statistique, comme le **test de Kruskal-Wallis**, de détecter une différence. Cela reflète une homogénéité totale des niveaux de

satisfaction au travail entre les groupes étudiés. Aucun diplôme ne semble donc influencer ces niveaux de satisfaction.

4 . Indépendance :

- ❖ La dépression est-elle indépendante des habitudes alimentaires (saines/modérées) ?

Le résultat du **test de Khi-Deux** (Stat = **1203.267**, p-value = **0.0000**) indique que la dépression n'est **pas indépendante** des habitudes alimentaires. La valeur p très faible (< 0.05) montre une relation statistiquement significative entre la dépression et le type d'habitudes alimentaires (saines ou modérées). Cela suggère que les habitudes alimentaires pourraient être un facteur associé à l'apparition de la dépression ou en être influencées. Une analyse plus approfondie pourrait explorer la nature précise de cette relation.

- ❖ La durée du sommeil (par exemple, moins de 5 heures, 5-6 heures, 7-8 heures) est-elle indépendante de la dépression ?

Le **test de Khi-Deux** (Stat = **277.135**, p-value = **0.0000**) montre que la durée du sommeil n'est **pas indépendante** de la dépression. La p-value extrêmement faible indique une relation statistiquement significative entre les deux variables. Cela suggère que la durée du sommeil pourrait être associée à la présence ou à l'absence de dépression. Les étudiants ayant des troubles de la dépression pourraient présenter des schémas de sommeil distincts, tels que des durées de sommeil réduites ou perturbées. Cette relation mérite d'être explorée davantage pour comprendre son impact sur la santé mentale.

Conclusion générale

L'analyse statistique menée sur les facteurs influençant la dépression chez les étudiants met en lumière une réalité préoccupante : un nombre significatif d'étudiants présente des symptômes dépressifs, souvent exacerbés par des facteurs académiques, économiques et sociaux.

Les données révèlent que des éléments tels que la pression académique, le stress financier, des habitudes de sommeil perturbées et une faible satisfaction dans les études ou le travail sont fortement associés à des niveaux accrus de dépression. Par exemple, les étudiants confrontés à une pression académique intense ou à des difficultés financières sont plus susceptibles de développer des symptômes dépressifs.

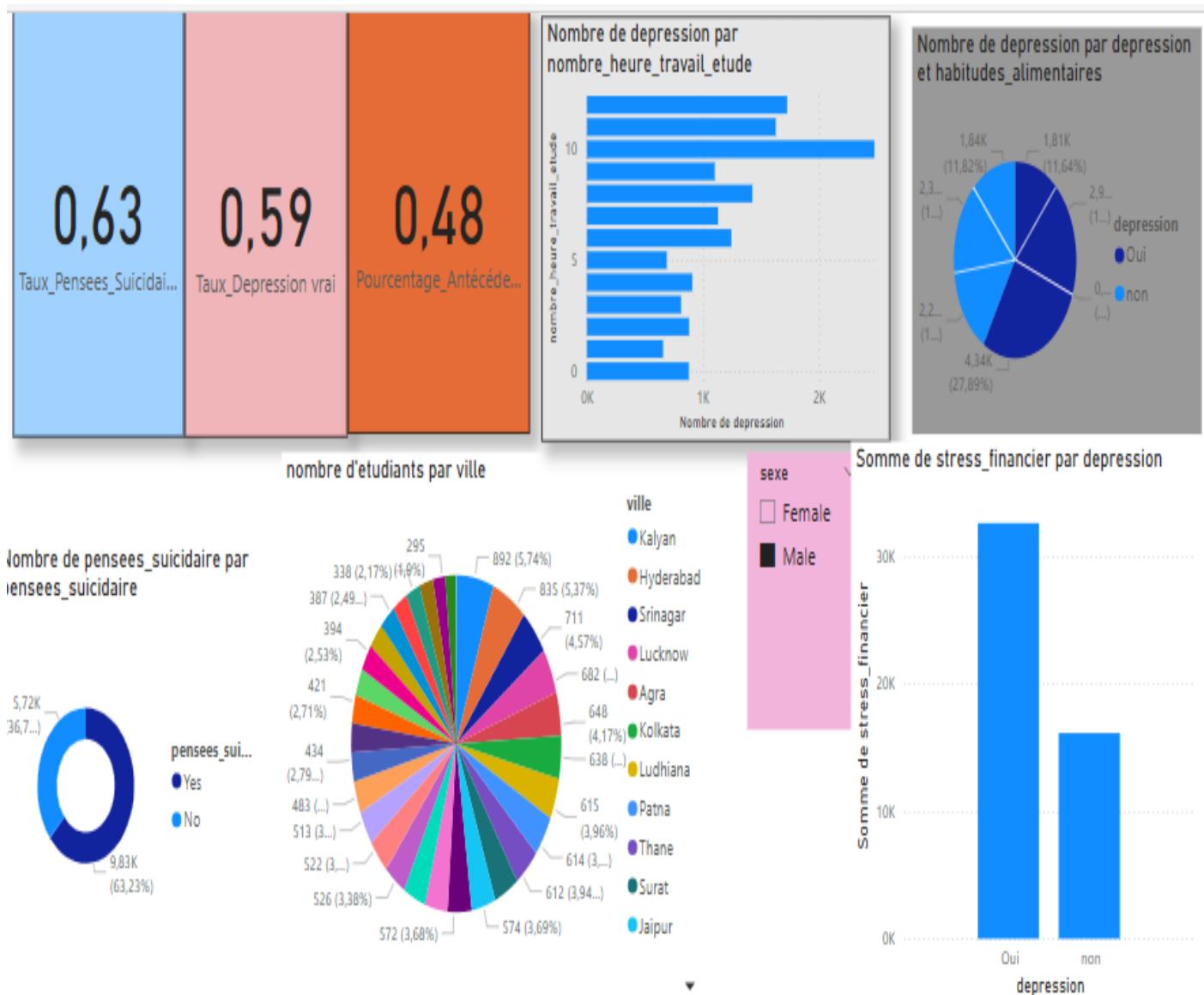
Ces résultats sont cohérents avec les tendances observées dans la littérature récente. Des études ont montré que la pandémie de COVID-19 a exacerbé les problèmes de santé mentale chez les étudiants, avec une augmentation notable des symptômes dépressifs et des idées suicidaires. L'isolement social, la précarité économique et les changements dans les modes d'enseignement ont contribué à cette détérioration du bien-être mental.[Le Monde.fr](#)

Face à cette situation, il est impératif de mettre en place des stratégies d'intervention ciblées. Cela inclut le renforcement des services de soutien psychologique sur les campus, la promotion de modes de vie sains et équilibrés, ainsi que des programmes visant à réduire la pression académique et à améliorer la satisfaction des étudiants dans leurs parcours éducatifs et professionnels.

En conclusion, cette étude souligne l'importance d'une approche holistique pour aborder la santé mentale des étudiants, en tenant compte des multiples facteurs qui contribuent à la dépression. Des efforts concertés entre les institutions éducatives, les professionnels de la santé et les décideurs politiques sont nécessaires pour créer un environnement propice au bien-être mental des étudiants.

Annexe

Tableau de bord



Codes utilisés

```
# Charger le fichier CSV
file_path = r"C:\INSEEDS\PROJET\STAT DESC\Student_Depression.csv"
df = pd.read_csv(file_path)

# Suppression de la colonne "id"
df.drop(columns=['id'], inplace=True)

# Aperçu des premières lignes
# Transformation en catégorielle
cols_to_categorical = [
    'sexe', 'profession', 'diplome_suivi', 'pensees_suicidaire',
    'antecedants_familiaux_maladie_mentale', 'duree_sommeil', 'habitudes_alimentaires'
```

```
]  
df[cols_to_categorical] = df[cols_to_categorical].astype('category')  
# Transformation de 'depression'  
df['depression'] = df['depression'].map({0: 'Non', 1: 'Oui'}).astype('category')  
  
# Transformation de l'âge en entier  
df['age'] = df['age'].astype(int)  
df.head()  
  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
# Calculer le nombre de valeurs manquantes pour chaque colonne  
missing_values = df.isnull().sum()  
  
# Filtrer uniquement les colonnes avec des valeurs manquantes  
missing_values = missing_values[missing_values > 0]  
  
# Créer un graphique en barres  
plt.figure(figsize=(10, 6))  
sns.barplot(x=missing_values.index, y=missing_values.values, palette="viridis")  
plt.title("Nombre de valeurs manquantes par colonne")  
plt.xlabel("Colonnes")  
plt.ylabel("Nombre de valeurs manquantes")  
plt.xticks(rotation=45) # Faire pivoter les étiquettes des colonnes  
plt.show()  
# Suppression ou imputation si <5%  
df = df.dropna(thresh=df.shape[0]*0.95, axis=1) # on garde si >=95% de valeurs présentes  
df = df.dropna() # ou bien .fillna(...) selon la logique métier  
  
# Visualisation après traitement  
plt.figure(figsize=(12,6))  
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')  
plt.title("Valeurs manquantes - Après traitement")  
plt.show()
```

```
# Supprimer les doublons
df = df.drop_duplicates()

from scipy.stats.mstats import winsorize

# Variables quantitatives
quant_vars = df.select_dtypes(include=['float64', 'int32']).columns

# Boxplot avant traitement
df[quant_vars].plot(kind='box', subplots=True, layout=(len(quant_vars)//3+1, 3), figsize=(15,10), sharex=False)
plt.suptitle("Boxplot avant winzorisation")
plt.tight_layout()
plt.show()

# Winzorisation (à 5%)
for var in quant_vars:
    df[var] = winsorize(df[var], limits=[0.05, 0.05])

# Boxplot après traitement
df[quant_vars].plot(kind='box', subplots=True, layout=(len(quant_vars)//3+1, 3), figsize=(15,10), sharex=False)
plt.suptitle("Boxplot après winzorisation")
plt.tight_layout()

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import kurtosis, skew

# Sélection des variables quantitatives (automatique ou manuelle si besoin)
quanti_vars = df.select_dtypes(include=['float64', 'int32']).columns.tolist()
print("Variables quantitatives :", quanti_vars)

# Taille du graphique
plt.figure(figsize=(16, 12))

# Afficher les distributions
for i, col in enumerate(quanti_vars):
    plt.subplot(3, 3, i + 1)
    sns.histplot(df[col], kde=True, bins=30, color='skyblue')
    plt.title(f'Distribution de {col}'')
```

```

plt.xlabel("")
plt.ylabel("")

plt.tight_layout()
plt.show()

# Fonction pour calculer les stats demandées

def statistiques_perso(df, variables):
    data = []
    for var in variables:
        stats = {
            'Variable': var,
            'Moyenne': df[var].mean(),
            'Médiane': df[var].median(),
            'Min': df[var].min(),
            'Max': df[var].max(),
            'Écart-type': df[var].std(),
            'Variance': df[var].var(),
            'Asymétrie (Skew)': skew(df[var].dropna()),
            'Kurtosis': kurtosis(df[var].dropna())
        }
        data.append(stats)
    return pd.DataFrame(data)

# Résumé des stats

resume_stats = statistiques_perso(df, quanti_vars)
print(resume_stats)

df = df.drop(columns=['pression_lie_travail', 'satisfaction_travail'], errors='ignore')

# Quantitatives

quanti_vars = df.select_dtypes(include=['float64', 'int32']).columns.tolist()

# Qualitatives (inclut les colonnes de type object et category)

qual_vars = df.select_dtypes(include=['object', 'category']).columns.tolist()

from scipy.stats import kruskal

import seaborn as sns

import matplotlib.pyplot as plt

for var in quanti_vars: # Parcourir les variables quantitatives
    plt.figure(figsize=(6, 4))
    sns.boxplot(x=df["depression"], y=df[var], palette="Set2")

```

```
plt.title(f"Boxplot : {var} en fonction de la dépression")
plt.xlabel("Dépression")
plt.ylabel(var)
plt.tight_layout()
plt.show()

# Suppression des valeurs manquantes
groups = [group.dropna() for name, group in df.groupby("depression")[var]]

# Test de Kruskal-Wallis
print(f"\n--- Test de Kruskal-Wallis entre {var} et depression ---")
try:
    stat, p = kruskal(*groups)
    print(f"Statistique = {stat:.3f}, p-value = {p:.4f}")

# Interprétation des résultats
if p < 0.05:
    print(f"➡ Relation significative entre {var} et depression.")
else:
    print(f"➡ Pas de relation significative entre {var} et depression.")

except Exception as e:
    print(f"Erreur : {e}")

import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency, fisher_exact

# Variables qualitatives pour la liaison avec la dépression
qual_vars = ['sexe', 'profession', 'diplome_suivi', 'pensees_suicidaire',
             'antecedants_familiaux_maladie_mentale', 'ville',
             'duree_sommeil', 'habitudes_alimentaires']

# Boucle pour tester chaque variable qualitative avec la dépression
results = [] # Pour stocker les résultats dans un tableau final

for var in qual_vars:
    # Création de la table de contingence
    contingency_table = pd.crosstab(df['depression'], df[var])
```

```
# Calcul des effectifs théoriques
_, p, _, expected = chi2_contingency(contingency_table)
min_expected = expected.min() # Trouver le plus petit effectif théorique

print(f"\n--- Liaison entre {var} et depression ---")
print("Table de contingence :")
print(contingency_table)

if min_expected >= 5:
    # Condition de Cochran respectée : Test de Khi-Deux
    chi2, p, _, _ = chi2_contingency(contingency_table)
    test = "Khi-Deux"
else:
    # Condition de Cochran non respectée : Test de Fisher
    try:
        # Le test de Fisher est seulement applicable sur des tables 2x2
        if contingency_table.shape == (2, 2):
            odds_ratio, p = fisher_exact(contingency_table)
            test = "Fisher"
        else:
            raise ValueError("Le test de Fisher n'est applicable qu'aux tables 2x2")
    except Exception as e:
        print(f"Erreur avec le test de Fisher : {e}")
        continue # Passer à la variable suivante

# Résumé des résultats
print(f"Test utilisé : {test}")
print(f"p-value : {p:.4f}")
if p < 0.05:
    conclusion = " ➔ Relation significative entre la variable et depression."
else:
    conclusion = " ➔ Pas de relation significative entre la variable et depression."
print(conclusion)

# Stocker les résultats dans une liste
results.append([var, test, p, conclusion])

# Affichage des résultats sous forme de tableau
```

```
results_table = pd.DataFrame(results, columns=['Variable', 'Test', 'p-value', 'Conclusion'])

print("\n📊 Tableau récapitulatif des résultats :")
print(results_table)

import pandas as pd

from scipy.stats import chi2_contingency

# Liste des variables qualitatives
qual_vars = ['sexe', 'profession', 'diplome_suivi', 'pensees_suicidaire',
    'antecedants_familiaux_maladie_mentale', 'ville',
    'duree_sommeil', 'habitudes_alimentaires']

# Liste pour stocker les résultats
results = []

# Boucle à travers les variables qualitatives
for var in qual_vars:
    # Création de la table de contingence
    contingency_table = pd.crosstab(df['depression'], df[var])

    # Calcul des effectifs théoriques
    _, _, _, expected = chi2_contingency(contingency_table)

    # Vérification de la condition de Cochran
    min_effectifs = expected.min() # Effectif théorique minimal
    cochrان_respecte = min_effectifs >= 5 # Condition respectée ou non

    # Stocker les résultats
    results.append({
        "Variable": var,
        "Effectifs théoriques minimaux": min_effectifs,
        "Condition de Cochran": "✅ Respectée" if cochrان_respecte else "❌ Non respectée"
    })

# Convertir les résultats en DataFrame pour affichage clair
results_df = pd.DataFrame(results)

# Affichage des résultats
print("\n📊 Résultats du calcul des effectifs théoriques :")
```

```
print(results_df)

import pandas as pd

from scipy.stats import chi2_contingency, fisher_exact

# Tableau des résultats des effectifs théoriques

results_df = pd.DataFrame({ 

    "Variable": ["sexe", "profession", "diplome_suivi", "pensees_suicidaire",
                 "antecedants_familiaux_maladie_mentale", "ville",
                 "duree_sommeil", "habitudes_alimentaires"],

    "Effectifs théoriques minimaux": [5119.584773, 0.414474, 14.506595, 4245.044304,
                                      5595.815578, 0.414474, 7.460535, 4.973690],

    "Condition de Cochran": [" Respectée", " Non respectée", " Respectée", " Respectée",
                             " Respectée", " Non respectée", " Respectée", " Non respectée"] 

})

# Liste pour stocker les résultats des tests

results_fisher = []
results_chi2 = []

# Boucle à travers les variables du tableau

for _, row in results_df.iterrows():

    var = row["Variable"]
    cochran = row["Condition de Cochran"]

    # Créer une table de contingence entre "depression" et la variable
    contingency_table = pd.crosstab(df['depression'], df[var])

    if cochran == " Non respectée":

        try:

            # Test de Fisher (applicable uniquement sur des tables 2x2)
            if contingency_table.shape == (2, 2):

                odds_ratio, p = fisher_exact(contingency_table)
                results_fisher.append([var, p, "Fisher"])

            else:

                results_fisher.append([var, None, "Fisher (non applicable à des tables > 2x2)"])

        except Exception as e:
            results_fisher.append([var, None, f"Erreur Fisher : {e}"])

    else:
        results_fisher.append([var, None, "Non applicable"])

else:
```

try:

```
# Test de Khi-Deux  
chi2, p, _, _ = chi2_contingency(contingency_table)  
results_chi2.append([var, p, "Khi-Deux"])  
  
except Exception as e:  
    results_chi2.append([var, None, f"Erreur Khi-Deux : {e}"])
```

Convertir les résultats en DataFrame pour un affichage clair

```
fisher_table = pd.DataFrame(results_fisher, columns=["Variable", "p-value", "Test"])  
chi2_table = pd.DataFrame(results_chi2, columns=["Variable", "p-value", "Test"])
```

Afficher les résultats

```
print("\n📊 Résultats des tests de Fisher (condition de Cochran non respectée) :")  
print(fisher_table)
```

```
print("\n📊 Résultats des tests de Khi-Deux (condition de Cochran respectée) :")
```

```
print(chi2_table)  
  
import pandas as pd  
  
from scipy.stats import chi2_contingency
```

Variables pour lesquelles Fisher n'était pas applicable

```
variables_to_test = ["profession", "ville", "habitudes_alimentaires"]
```

Liste pour stocker les résultats

```
results_chi2 = []
```

for var in variables_to_test:

Création de la table de contingence

```
contingency_table = pd.crosstab(df['depression'], df[var])
```

Application du test de Khi-Deux

try:

```
chi2, p, _, _ = chi2_contingency(contingency_table)
```

```
results_chi2.append([var, chi2, p, "Khi-Deux"])
```

except Exception as e:

```
results_chi2.append([var, None, None, f"Erreur Khi-Deux : {e}"])
```

Affichage des résultats sous forme de tableau

```
chi2_table = pd.DataFrame(results_chi2, columns=["Variable", "Chi2-stat", "p-value", "Test"])
```

```
print("\n📊 Résultats des tests de Khi-Deux :")  
print(chi2_table)  
import pandas as pd  
import matplotlib.pyplot as plt  
  
# Liste des variables qualitatives  
qual_vars = ['sexe', 'profession', 'diplome_suivi', 'pensees_suicidaire',  
             'antecedants_familiaux_maladie_mentale', 'ville',  
             'duree_sommeil', 'habitudes_alimentaires']  
  
# Boucle pour créer un diagramme en barres pour chaque variable  
for var in qual_vars:  
    # Table de contingence entre depression et la variable  
    contingency_table = pd.crosstab(df['depression'], df[var])  
  
    # Création du graphique en barres  
    contingency_table.plot(kind='bar', stacked=True, figsize=(8, 6), colormap='viridis')  
    plt.title(f"Relation entre depression et {var}")  
    plt.xlabel("Dépression")  
    plt.ylabel("Nombre d'observations")  
    plt.xticks(rotation=0)  
    plt.tight_layout()  
    plt.show()  
  
from statsmodels.stats.proportion import proportion_confint  
  
# Nombre d'étudiants ayant des pensées suicidaires  
n_positive = df[df['pensees_suicidaire'] == 'Yes'].shape[0]  
  
# Taille totale de l'échantillon  
n_total = df.shape[0]  
  
# Intervalle de confiance (méthode de Wilson)  
conf_int = proportion_confint(count=n_positive, nobs=n_total, method='wilson')  
print(f"Intervalle de confiance pour la proportion : {conf_int[0]:.4f} - {conf_int[1]:.4f}")  
# Filtrer les étudiants souffrant de dépression  
depression_filter = df[df['depression'] == 'Oui']  
# Moyenne et médiane  
mean_hours = depression_filter['nombre_heure_travail_etude'].mean()
```

```
median_hours = depression_filter['nombre_heure_travail_etude'].median()

print(f"Moyenne des heures de travail/étude : {mean_hours:.2f}")
print(f"Médiane des heures de travail/étude : {median_hours:.2f}")

# ---

# Moyenne et médiane pour les étudiants avec dépression
mean_stress_dep = depression_filter['stress_financier'].mean()
median_stress_dep = depression_filter['stress_financier'].median()

# Moyenne et médiane pour les étudiants sans dépression
non_depression_filter = df[df['depression'] == 'Non']
mean_stress_non = non_depression_filter['stress_financier'].mean()
median_stress_non = non_depression_filter['stress_financier'].median()

print("Avec dépression - Moyenne : ", mean_stress_dep, ", Médiane : ", median_stress_dep)
print("Sans dépression - Moyenne : ", mean_stress_non, ", Médiane : ", median_stress_non)

# ---

import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import mannwhitneyu, ttest_ind

# Étape 1 : Visualisation
sns.boxplot(x='depression', y='satisfaction_etudes', data=df)
plt.title('Satisfaction des études vs Dépression')
plt.show()

# Étape 2 : Statistiques de base
print(df.groupby('depression')['satisfaction_etudes'].describe())

# Étape 3 : Tester la normalité
from scipy.stats import shapiro

for group in df['depression'].unique():
    stat, p = shapiro(df[df['depression'] == group]['satisfaction_etudes'])
```

```

print(f"Normalité pour {group} : Stat = {stat:.3f}, p = {p:.4f}")

# Étape 4 : Comparer les moyennes (Test de Mann-Whitney si non normal)
stat, p = mannwhitneyu(depression_filter['satisfaction_etudes'], non_depression_filter['satisfaction_etudes'])
print(f"Test de Mann-Whitney : Stat = {stat:.3f}, p-value = {p:.4f}")

# Étape 1 : Visualisation
sns.boxplot(x='diplome_suivi', y=' satisfaction_travail', data=df)
plt.title('Satisfaction au travail vs Diplôme suivi')
plt.xticks(rotation=45)
plt.show()

# Étape 2 : Statistiques de base
print(df.groupby('diplome_suivi')[' satisfaction_travail'].describe())

# Étape 3 : Tester la normalité pour chaque groupe
for diploma in df['diplome_suivi'].unique():
    stat, p = shapiro(df[df['diplome_suivi'] == diploma][' satisfaction_travail'])
    print(f"Normalité pour {diploma} : Stat = {stat:.3f}, p = {p:.4f}")

# Étape 4 : Test de Kruskal-Wallis si non normal
from scipy.stats import kruskal

groups = [df[df['diplome_suivi'] == diploma][' satisfaction_travail'].dropna() for diploma in df['diplome_suivi'].unique()]
stat, p = kruskal(*groups)
print(f"Test de Kruskal-Wallis : Stat = {stat:.3f}, p-value = {p:.4f}")

# Test d'indépendance (Khi-Deux)
contingency_table = pd.crosstab(df['duree_sommeil'], df['depression'])
chi2, p, _, _ = chi2_contingency(contingency_table)
print(f"Test de Khi-Deux : Stat = {chi2:.3f}, p-value = {p:.4f}")

# Test d'indépendance (Khi-Deux)
contingency_table = pd.crosstab(df['depression'], df['habitudes_alimentaires'])
chi2, p, _, _ = chi2_contingency(contingency_table)
print(f"Test de Khi-Deux : Stat = {chi2:.3f}, p-value = {p:.4f}")

.

for col in quant_vars:
    df[col] = pd.to_numeric(df[col], errors='coerce')
import pandas as pd

```

```
import matplotlib.pyplot as plt
from prince import MCA
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from scipy.cluster.hierarchy import dendrogram, linkage

# -----
# ANALYSE DES CORRESPONDANCES MULTIPLES (ACM)
# -----

def analyse_acm(df, qual_vars):
    print("\n--- ANALYSE DES CORRESPONDANCES MULTIPLES (ACM) ---")

    # Vérification et conversion des variables qualitatives
    df[qual_vars] = df[qual_vars].astype(str)
    df = df.dropna(subset=qual_vars)

    # Application de l'ACM
    mca = MCA(n_components=2, random_state=42)
    mca_result = mca.fit(df[qual_vars])

    # Graphe des modalités
    mod_coords = mca.column_coordinates(df[qual_vars])
    plt.figure(figsize=(8, 8))
    plt.scatter(mod_coords[0], mod_coords[1], alpha=0.7, color="purple")
    for i, text in enumerate(mod_coords.index):
        plt.text(mod_coords[0][i], mod_coords[1][i], text, color="black")
    plt.axhline(0, color="gray", linestyle="--")
    plt.axvline(0, color="gray", linestyle="--")
    plt.title("Graphe des modalités (ACM)")
    plt.xlabel("Dimension 1")
    plt.ylabel("Dimension 2")
    plt.show()

# -----
# ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC)
# -----

def analyse_afc(df, qual_vars):
```

```
print("\n--- ANALYSE FACTORIELLE DES CORRESPONDANCES (AFC) ---")

# Création d'un tableau croisé
afc_table = pd.crosstab(df[qual_vars[0]], df[qual_vars[1]])

# Calcul des fréquences relatives
afc_table_rel = afc_table / afc_table.sum().sum()

# Application d'une AFC avec prince MCA (alternative pour AFC)
mca = MCA(n_components=2, random_state=42)
mca_result = mca.fit(afc_table_rel)

# Graphe des correspondances
row_coords = mca.row_coordinates(afc_table_rel)
plt.figure(figsize=(8, 8))
plt.scatter(row_coords[0], row_coords[1], alpha=0.7, color="blue")
for i, text in enumerate(row_coords.index):
    plt.text(row_coords[0][i], row_coords[1][i], text, color="black")
plt.axhline(0, color="gray", linestyle="--")
plt.axvline(0, color="gray", linestyle="--")
plt.title("Graphe des correspondances (AFC)")
plt.xlabel("Dimension 1")
plt.ylabel("Dimension 2")
plt.show()

# -----
# CLUSTERING : Hiérarchique et K-Means
# -----

def clustering(df, qual_vars):
    print("\n--- CLUSTERING DES VARIABLES QUALITATIVES ---")

    # Conversion en données numériques pour clustering
    df_encoded = pd.get_dummies(df[qual_vars], drop_first=True)
    scaler = StandardScaler()
    df_scaled = scaler.fit_transform(df_encoded)

    # Clustering Hiérarchique
    linked = linkage(df_scaled, method="ward")
```

```
plt.figure(figsize=(10, 7))

dendrogram(linked, truncate_mode="lastp", p=30, leaf_rotation=45, leaf_font_size=10)
plt.title("Dendrogramme (Clustering Hiérarchique)")
plt.show()

# Clustering K-Means
silhouette_scores = []
k_values = range(2, 10)
for k in k_values:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(df_scaled)
    score = silhouette_score(df_scaled, kmeans.labels_)
    silhouette_scores.append(score)

# Optimisation du nombre de clusters
optimal_k = silhouette_scores.index(max(silhouette_scores)) + 2
print(f"Nombre optimal de clusters : {optimal_k}")

# Application de K-Means
kmeans = KMeans(n_clusters=optimal_k, random_state=42)
clusters = kmeans.fit_predict(df_scaled)
df["Cluster"] = clusters

# Résumé des clusters
print("\nRésumé des clusters :")
print(df.groupby("Cluster").mean())

# -----
# APPLICATION DU CODE AUX VARIABLES : profession, ville, habitudes_alimentaires
# -----
qual_vars = ['profession', 'ville', 'habitudes_alimentaires']

# ACM
analyse_acm(df, qual_vars)

# AFC (utilisant deux variables à la fois pour un tableau croisé)
analyse_afc(df, qual_vars[:2]) # Exemple avec "profession" et "ville"
```

```
df = df.dropna(subset=['depression', 'ville', 'profession', 'habitudes_alimentaires'])

df[['depression', 'ville', 'profession', 'habitudes_alimentaires']] = df[['depression', 'ville', 'profession',
'habitudes_alimentaires']].astype(str)

# 📦 IMPORTS
# =====

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.decomposition import PCA

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans

from scipy.cluster.hierarchy import dendrogram, linkage

from sklearn.metrics import silhouette_score

import prince

# =====

# 📈 CHARGEMENT DES DONNÉES
# =====

# Remplace "fichier.csv" par le nom de ton fichier
# df = pd.read_csv("fichier.csv")

# Détection des types de variables
qual_vars = df.select_dtypes(include=['object', 'category']).columns
quant_vars = df.select_dtypes(include=['float64', 'int64']).columns

# =====

# 🎨 ACP - Méthodes Factorielles (quantitatives)
# =====

def analyse_acp(df, quant_vars):
    print("\n⌚ Analyse en Composantes Principales (ACP)")

    df_clean = df.dropna(subset=quant_vars)
    X_scaled = StandardScaler().fit_transform(df_clean[quant_vars])

    pca = PCA()
    X_pca = pca.fit_transform(X_scaled)
```

```

# 1.1 - Graphique des valeurs propres
plt.figure(figsize=(8,5))
plt.plot(np.cumsum(pca.explained_variance_ratio_)*100, marker='o', linestyle='--')
plt.title("Graphique des valeurs propres (Variance expliquée cumulée)")
plt.xlabel("Nombre de composantes")
plt.ylabel("% de variance expliquée")
plt.grid(True)
plt.show()

# 1.2 - Interprétation des inerties
for i, var in enumerate(pca.explained_variance_ratio_[:5]):
    print(f"Composante {i+1} : {var*100:.2f}% de la variance expliquée")

# 1.3 - Cercle de corrélation
pcs = pca.components_[:,2]
plt.figure(figsize=(8,8))
for i, var in enumerate(quant_vars):
    plt.arrow(0, 0, pcs[0,i], pcs[1,i], color='r', alpha=0.5)
    plt.text(pcs[0,i]*1.1, pcs[1,i]*1.1, var, fontsize=12)
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("Cercle de corrélation")
plt.axhline(0, color='grey')
plt.axvline(0, color='grey')
plt.grid()
plt.show()

return X_pca, pca

X_pca, pca_model = analyse_acp(df, quant_vars)

# =====
#  ACM - Variables qualitatives
# =====

def analyse_acm(df, qual_vars):
    print("\n<img alt='ABC icon' style='vertical-align: middle;"/> Analyse des Correspondances Multiples (ACM)")



```

```

df_acm = df.dropna(subset=qual_vars)
mca = prince.MCA(n_components=2, random_state=42)
mca = mca.fit(df_acm[qual_vars])

# 1.4 - Graphe des modalités
mod_coords = mca.column_coordinates(df_acm[qual_vars])
plt.figure(figsize=(8,8))
plt.scatter(mod_coords[0], mod_coords[1], color='purple')
for i, txt in enumerate(mod_coords.index):
    plt.text(mod_coords[0][i], mod_coords[1][i], txt, fontsize=10)
plt.axhline(0, color='grey', linestyle='--')
plt.axvline(0, color='grey', linestyle='--')
plt.title("Graphe des modalités (ACM)")
plt.xlabel("Dimension 1")
plt.ylabel("Dimension 2")
plt.grid()
plt.show()

```

```
analyse_acm(df, qual_vars)
```

```
# =====
# 🤖 CLUSTERING
# =====
```

```

# 2.1 Dendrogramme (Clustering hiérarchique)
def dendrogramme(df, quant_vars):
    print("\n📊 Dendrogramme")
    df_clean = df.dropna(subset=quant_vars)
    X = StandardScaler().fit_transform(df_clean[quant_vars])
    linkage_matrix = linkage(X, method='ward')
    plt.figure(figsize=(10, 6))
    dendrogram(linkage_matrix, truncate_mode='lastp', p=20)
    plt.title("Dendrogramme")
    plt.xlabel("Observations")
    plt.ylabel("Distance")
    plt.show()

```

```
dendrogramme(df, quant_vars)
```

```
# 2.2 Nombre optimal de clusters avec Silhouette

def optimal_clusters(df, quant_vars):

    print("\n🔍 Détermination du nombre optimal de clusters")
    df_clean = df.dropna(subset=quant_vars)

    X = StandardScaler().fit_transform(df_clean[quant_vars])

    scores = []

    for k in range(2, 10):

        kmeans = KMeans(n_clusters=k, random_state=42)

        labels = kmeans.fit_predict(X)

        score = silhouette_score(X, labels)

        scores.append(score)

    plt.plot(range(2,10), scores, marker='o', color='orange')

    plt.title("Score Silhouette")

    plt.xlabel("Nombre de clusters")

    plt.ylabel("Silhouette")

    plt.grid(True)

    plt.show()

    best_k = np.argmax(scores) + 2

    print(f"✅ Meilleur nombre de clusters : {best_k}")

    return best_k
```

```
k = optimal_clusters(df, quant_vars)
```

```
# 2.3 Caractérisation des clusters

def caracteriser_clusters(df, quant_vars, k):

    print("\n💬 Caractérisation des clusters")

    df_clean = df.dropna(subset=quant_vars).copy()

    X = StandardScaler().fit_transform(df_clean[quant_vars])

    kmeans = KMeans(n_clusters=k, random_state=42)

    df_clean['Cluster'] = kmeans.fit_predict(X)
```

```
# Moyennes par cluster

print("\n📌 Moyennes des variables par cluster :")

print(df_clean.groupby('Cluster')[quant_vars].mean())
```

```
# 2D visualisation sur ACP
```

```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)
plt.figure(figsize=(10,7))
sns.scatterplot(x=X_pca[:,0], y=X_pca[:,1], hue=df_clean['Cluster'], palette='tab10')
plt.title("Clusters projetés sur les axes de l'ACP")
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.legend(title="Cluster")
plt.grid(True)
plt.show()

characteriser_clusters(df, quant_vars, k)

# ---
```

