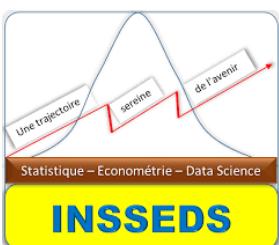


MINISTERE DE L'ENSEIGNEMENT SUPERIEUR  
ET DE LA RECHERCHE SCIENTIQUE

REPUBLIQUE DE CÔTE D'IVOIRE



INSTITUT SUPERIEUR DE  
STATISTIQUE D'ECONOMETRIE  
ET DATA SCIENCE



UNION-DISCIPLINE-TRAVIL

## STATISTIQUE INFÉRENTIELLE

Niveau : Master 1

**Étude des facteurs influençant  
la dépression chez les étudiants  
à l'aide de méthodes statistiques**

Année Universitaire : 2024-2025

Nom : DAGNOGO

Prénoms : CALIGNINRI SAFIATOU

Nom de l'enseignant :  
**Akposso Didier Martial**

## Avant-propos

Ce mini-projet s'inscrit dans le cadre d'une analyse inférentielle visant à étudier les facteurs influençant la dépression chez les étudiants. La santé mentale des jeunes en milieu académique est un enjeu majeur, particulièrement dans un contexte où les pressions académiques, financières et sociales s'intensifient. À travers des méthodes statistiques rigoureuses et des techniques d'analyse exploratoire, ce projet permettra de **mettre en lumière les relations existantes entre le mode de vie, les habitudes académiques et la santé mentale**. Les résultats obtenus permettront de mieux comprendre les mécanismes sous-jacents à la détresse psychologique des étudiants et pourront servir de base à des recommandations pour des politiques de prévention et d'accompagnement adaptées.

## Table des matières

<b>Avant-propos .....</b>	2
<b>Introduction Générale.....</b>	6
<b>I) PRETRATEMENT DES DONNEES .....</b>	7
1) Visualisation des valeurs manquantes .....	8
2) Traitement des valeurs manquantes .....	8
3) Visualisation des valeurs aberrantes .....	9
4) Traitement des valeurs aberrantes.....	9
<b>II) ANALYSE UNIVARIEE .....</b>	10
1) Analyse univariée des variables numériques .....	10
1.1) Résumés statistiques .....	10
1.2) Représentations graphiques .....	11
2) Analyse univariée des variables catégorielles.....	12
2.1) Proportions et Interprétations.....	12
2.2) Représentations graphiques.....	13
<b>Visualisation spéciale pour les variables clés.....</b>	14
<b>III) ANALYSE BIVARIEE .....</b>	16
<b>Normalité .....</b>	16
1) Pour une variable quantitative et qualitative.....	17
2) Pour deux variables qualitatives .....	21
<b>Condition de Cochran .....</b>	21
<b>IV) STATISTIQUE INFÉRENTIELLE .....</b>	26
1) Trouvons l'intervalle de confiance pour la proportion d'étudiants ayant déjà eu des pensées suicidaires .....	26
2) Estimation de la moyenne et la médiane des heures de travail ou d'études pour les étudiants souffrant de dépression .....	26
3) Évaluons la moyenne et la médiane du stress financier pour les étudiants avec et sans dépression.....	26
4) La satisfaction des études diffère-t-elle significativement entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas ? .....	27
5) Les niveaux de satisfaction au travail diffèrent-ils significativement selon le diplôme suivi ? .....	27
6) La dépression est-elle indépendante des habitudes alimentaires (saines/modérées) ? .....	27
7) La durée du sommeil est-elle indépendante de la dépression ? .....	28
<b>Conclusion .....</b>	30
<b>TABLEAU DE BORD .....</b>	31

**ANNEXE .....**.....31**Liste des tableaux***Tableau 1: Extrait du jeu de donnée**Tableau 2: Résumé statistique**Tableau 3 : Tableau des proportions des variables catégorielles**Tableau 4: Test statistique de normalité**Tableau 5: Table de vérification de la condition de Cochran***Liste des figures***Figure 1: Visualisation des valeurs manquantes**Figure 2: Visualisation des valeurs manquantes après traitement**Figure 3: Visualisation des valeurs aberrantes des variables quantitatives**Figure 4: Visualisation des valeurs aberrantes après traitement**Figure 5: Représentations graphique des variables numériques**Figure 6: Représentations graphique des variables catégorielles**Figure 7: Répartition des variables catégorielles clés**Figure 8: Boxplots des variables numériques clés**Figure 9: Test graphique de normalité**Figure 10: Boxplot de l'âge par statut dépressifs**Figure 11: Boxplot pression\_académique par dépressifs**Figure 12: Boxplot pression\_liee\_travail par dépressifs**Figure 13: Boxplot de satisfaction travail par dépressifs**Figure 13: Boxplot moyenne\_notes par dépressifs**Figure 14: Boxplot satisfaction étude par dépressifs**Figure 15: Boxplot du nombre\_heure\_travail\_etude par dépressifs**Figure 16: Boxplot du stress financier par dépressifs**Figure 17: Distribution du sexe par dépressifs**Figure 18: Distribution de la ville par dépressifs**Figure 19: Distribution de la profession et du statut deprssif**Figure 20: Distribution de la durée de sommeil par dépressifs**Figure 21: Distribution des habitudes alimentaires par dépressifs**Figure 22: Distribution du diplôme suivi par dépressifs**Figure 23: Distribution de la pensée suicidaire par dépressifs*

*Figure 24: Distribution des antécédants familiaux de maladie mentale par dépressifs*

*Figure 25: Répartition des cas de dépression par habitudes alimentaires*

*Figure 26: Répartition de la dépression en fonction de la durée de sommeil*

# Introduction Générale

## 📌 Contexte et justification de l'étude

Dans un monde académique en constante évolution, où la pression scolaire et les exigences professionnelles s'intensifient, la santé mentale des étudiants est un enjeu crucial. Divers facteurs, tels que le stress académique, le manque de sommeil, les difficultés financières et les antécédents familiaux, jouent un rôle fondamental dans le développement de troubles dépressifs. L'objectif de cette étude est donc de quantifier et analyser l'impact de ces facteurs sur la prévalence de la dépression chez les étudiants.

## ❓ Problématique

Quels sont les principaux facteurs associés à la dépression chez les étudiants ? Existe-t-il une relation significative entre le mode de vie, les performances académiques et la santé mentale ? Cette étude vise à répondre à ces questions en fournissant une analyse statistique détaillée permettant de mieux comprendre les éléments déclencheurs et les interactions entre différentes variables.

## 🔍 Principaux résultats attendus

- Identifier les variables les plus influentes dans le développement de la dépression.
- Quantifier l'impact du stress académique et financier sur la santé mentale.
- Explorer la relation entre la durée du sommeil, la satisfaction académique et la présence de symptômes dépressifs.
- Proposer des recommandations basées sur les résultats statistiques pour améliorer la prise en charge des étudiants à risque.

## 🛠️ Méthodologie

L'étude repose sur une approche quantitative, combinant des techniques de prétraitement et d'analyse statistique avancées :

- Techniques de prétraitement des données : Nettoyage, gestion des valeurs manquantes, transformation des variables qualitatives et normalisation des données numériques.
- Analyses statistiques :
  - Estimation des intervalles de confiance pour la proportion d'étudiants ayant eu des pensées suicidaires.
  - Analyse des moyennes et médianes pour évaluer les différences entre étudiants souffrant de dépression et ceux qui n'en souffrent pas.
  - Tests de différence de moyennes pour comparer la satisfaction académique et professionnelle selon les groupes.
  - Tests d'indépendance (Khi-2) pour vérifier si la dépression est liée aux habitudes alimentaires ou à la durée du sommeil.
- Outils utilisés : Les analyses seront réalisées avec Python et Excel, et un tableau de bord interactif sous Power BI sera conçu pour visualiser les résultats de manière dynamique.

Ce rapport fournira des résultats précis et exploitables, contribuant à une meilleure compréhension des dynamiques de la santé mentale étudiante et à l'élaboration de stratégies de prévention et d'accompagnement.

## Dictionnaire des données

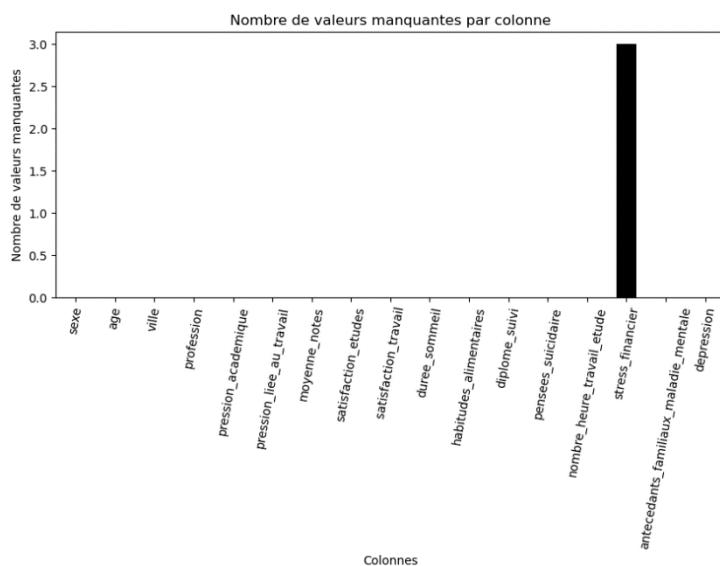
Variable	Description	Type initial	Type analytique
<b>sexe</b>	Sexe de l'étudiant (Masculin/Féminin)	object	category
<b>age</b>	Âge de l'étudiant	float64	float64
<b>ville</b>	Ville de résidence	object	category
<b>profession</b>	Statut professionnel (étudiant à temps plein ou autre activité)	object	category
<b>pression_academique</b>	Niveau de pression académique	float64	float64
<b>pression_liee_au_travail</b>	Niveau de pression liée au travail	float64	float64
<b>moyenne_notes</b>	Moyenne générale des notes	float64	float64
<b>satisfaction_etudes</b>	Niveau de satisfaction par rapport aux études	float64	float64
<b>satisfaction_travail</b>	Niveau de satisfaction par rapport au travail	float64	float64
<b>duree_sommeil</b>	Durée moyenne du sommeil (ex. : 5-6h, moins de 5h)	object	category
<b>habitudes_alimentaires</b>	Type d'habitudes alimentaires (saines, modérées)	object	category
<b>diplome_suivi</b>	Diplôme suivi ou obtenu (BSc, M.Tech, etc.)	object	category
<b>pensees_suicidaires</b>	A déjà eu des pensées suicidaires (Oui/Non)	object	category
<b>nombre_heure_travail_etude</b>	Nombre d'heures de travail ou d'études par jour	float64	float64
<b>stress_financier</b>	Score du stress financier	float64	float64
<b>antecedents_familiaux_maladies_mentales</b>	Antécédents familiaux de maladies mentales (Oui/Non)	object	category
<b>depression</b>	Présence d'un diagnostic de dépression (0 = Non, 1 = Oui)	int64	category

Tableau 6: Extrait du jeu de donnée

## I) PRETRATEMENT DES DONNEES

Avant d'effectuer les analyses statistiques, il est essentiel de préparer et nettoyer les données afin d'assurer leur fiabilité et leur cohérence. Cette étape de prétraitement permet d'éliminer les incohérences, de gérer les valeurs manquantes et de transformer les variables pour faciliter leur exploitation.

## 1) Visualisation des valeurs manquantes



L'ensemble des données contient trois (3) valeurs manquantes précisément au niveau de la colonnes stress\_financier. Nous constatons qu'il n'y a pas de doublons.

Figure 27: Visualisation des valeurs manquantes

## 2) Traitement des valeurs manquantes

La proportion de valeurs manquantes par rapport à la taille du jeu de donnée est moins de 5% alors, nous procèderons à la suppression de ces dites valeurs.

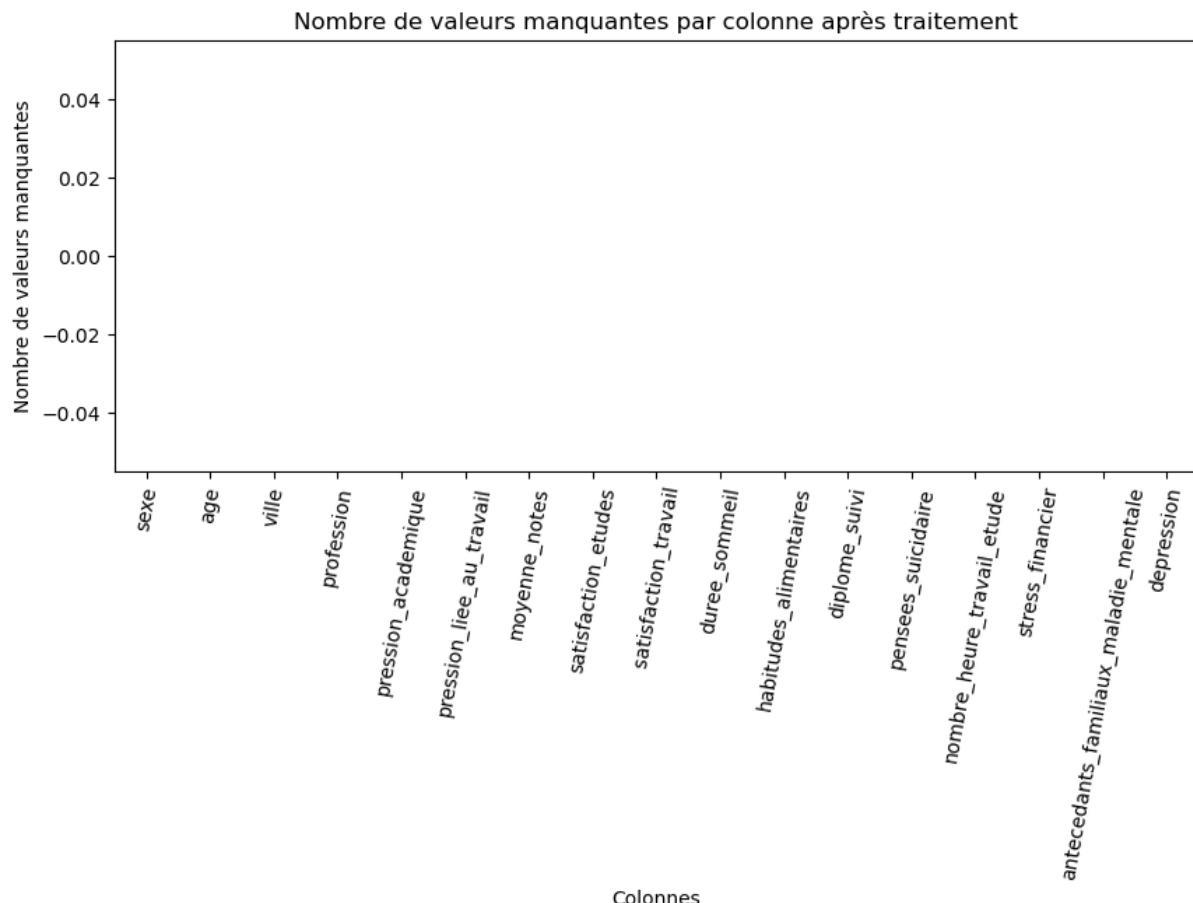


Figure 28: Visualisation des valeurs manquantes après traitement

### 3) Visualisation des valeurs aberrantes

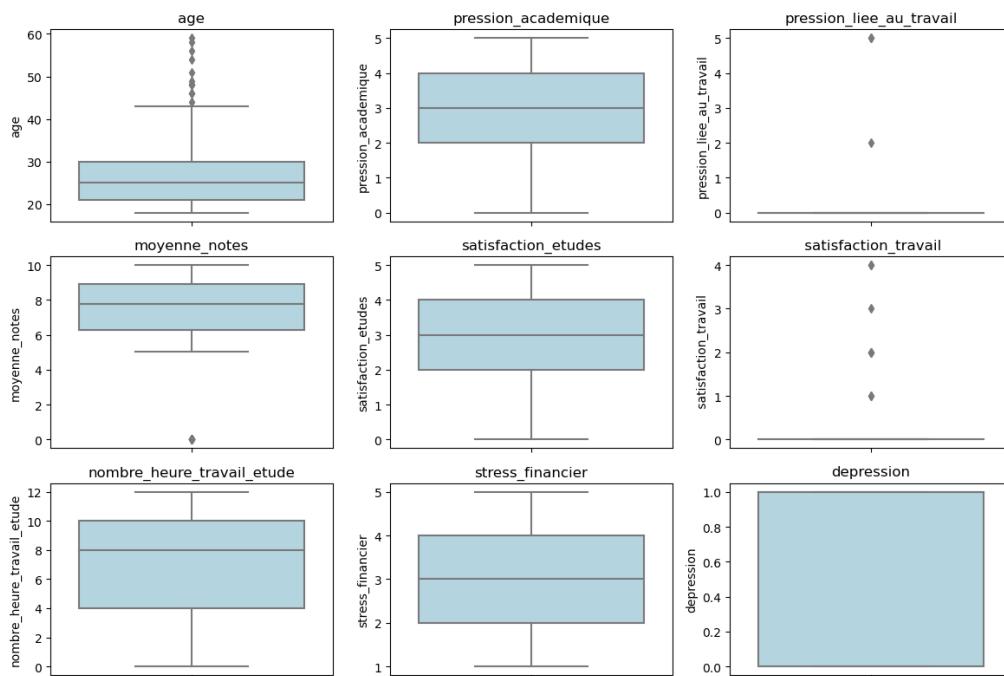


Figure 29: Visualisation des valeurs aberrantes des variables quantitatives

### 4) Traitement des valeurs aberrantes

Les valeurs aberrantes ont été traitées par la méthode de Winsorisation.

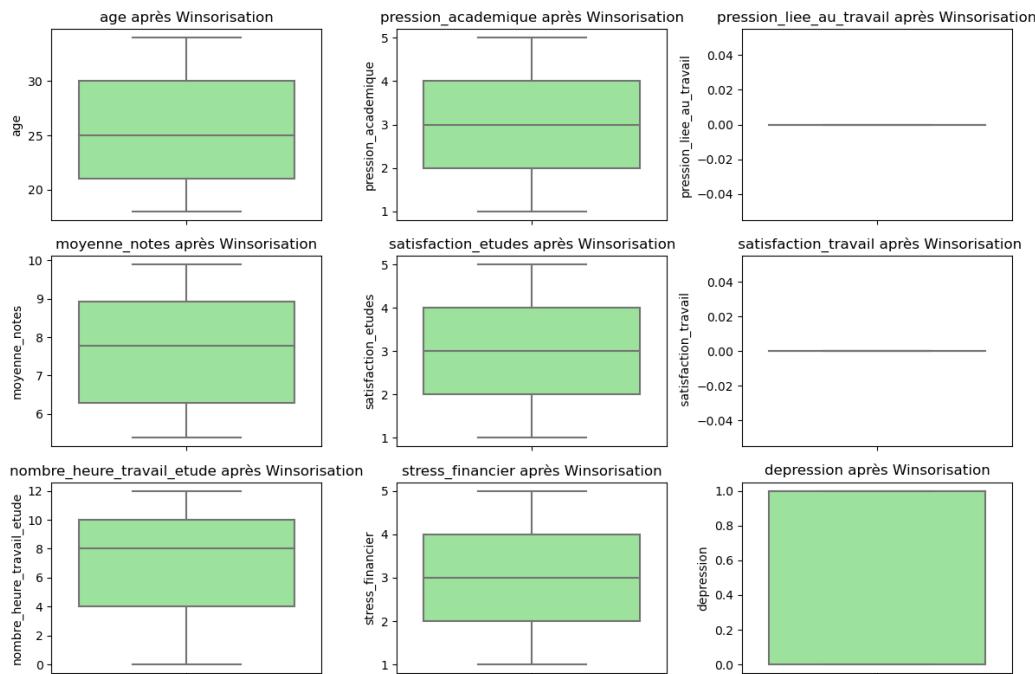


Figure 30: Visualisation des valeurs aberrantes après traitement

## II) ANALYSE UNIVARIEE

### 1) Analyse univariée des variables numériques

#### 1.1) Résumés statistiques

Variable	Min	Max	Moyenne	Médiane	Mode	Variance	Écart-type	Skewness	Kurtosis
age	18.0	59.0	25.82	25.00	24.00	24.07	4.91	0.13	-0.85
pression_academique	0.0	5.0	3.14	3.00	3.00	1.91	1.38	-0.14	-1.16
pression_liee_au_travail	0.0	5.0	0.00	0.00	0.00	0.00	0.04	108.58	12107.16
moyenne_notes	0.0	10.0	7.66	7.77	8.04	2.16	1.47	-0.11	-1.02
satisfaction_etudes	0.0	5.0	2.94	3.00	4.00	1.85	1.36	0.01	-1.22
satisfaction_travail	0.0	4.0	0.00	0.00	0.00	0.00	0.04	74.10	5925.30
nombre_heure_travail_etude	0.0	12.0	7.16	8.00	10.00	13.75	3.71	-0.45	-1.00
stress_financier	1.0	5.0	3.14	3.00	5.00	2.07	1.44	-0.13	-1.32

Tableau 7: Résumé statistique

#### Interprétation des résultats

1. **age** : La distribution est presque symétrique (skew = 0.13) et modérément étalée (écart-type = 4.91). Cela suggère une population jeune et homogène.
2. **pression\_academique** : Moyenne proche de 3 avec une légère asymétrie gauche. L'échelle de pression est modérée pour la plupart des étudiants.
3. **pression\_liee\_au\_travail** : Tous les étudiants semblent avoir répondu "0" à cette question. Le skew et la kurtosis sont extrêmement élevés car la distribution est ultra-concentrée en 0.
4. **moyenne\_notes** : Moyenne élevée (7.66) avec une légère asymétrie gauche. Cela montre que les étudiants ont de bonnes performances académiques.
5. **satisfaction\_etudes** : Moyenne quasi neutre (2.94) avec une distribution très plate (kurtosis = -1.22). La satisfaction est très variée, sans tendance marquée.
6. **satisfaction\_travail** : Valeurs quasi nulles avec des indices de forme extrêmes → les étudiants n'ont pas ou peu noté cette variable, probablement car ils ne travaillent pas.
7. **nombre\_heure\_travail\_etude** : Moyenne de 7.16 h/j, légère asymétrie à gauche. Les étudiants passent en moyenne une bonne partie de la journée à étudier.
8. **stress\_financier** : Moyenne modérée (3.14), distribution légèrement asymétrique à gauche. Le stress financier touche beaucoup d'étudiants, de façon diverse.

## 1.2) Représentations graphiques

Analyse univariée des variables quantitatives

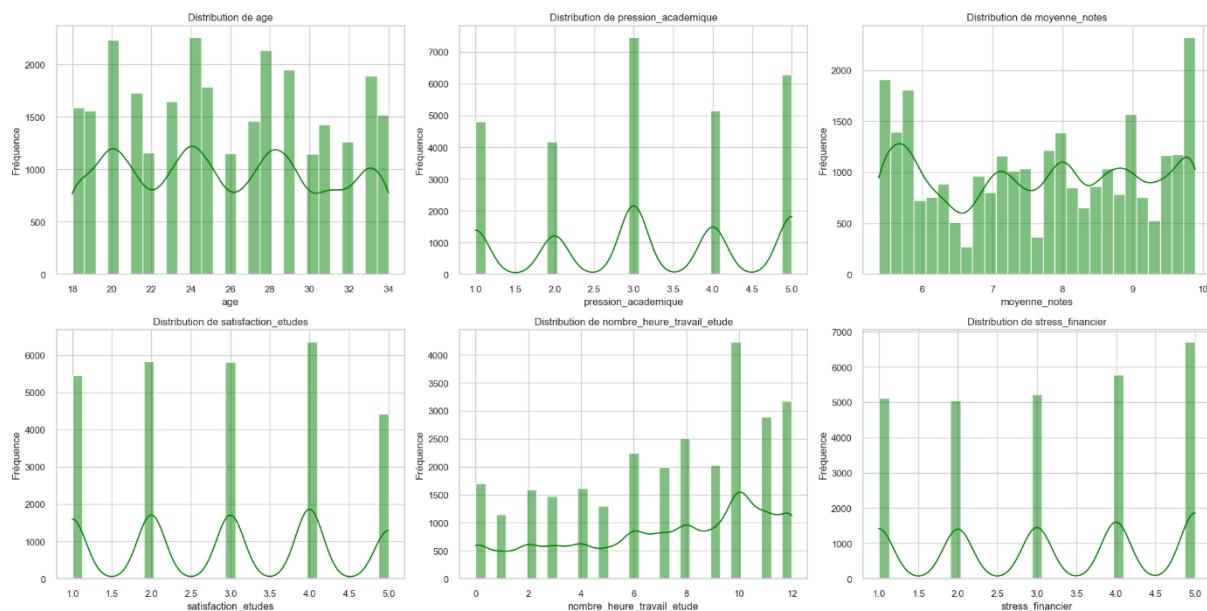


Figure 31: Représentations graphique des variables numériques

### Interprétation des distributions des variables étudiées

Ce graphique présente **six histogrammes** illustrant la répartition des variables **âge, pression académique, moyenne des notes, satisfaction aux études, nombre d'heures d'étude et stress financier**. Voici les observations principales :

#### Distribution de l'âge

La majorité des étudiants ont **entre 20 et 30 ans**, ce qui reflète une population étudiante classique.

La distribution semble **symétrique**, suggérant une répartition homogène de l'âge.

#### Distribution de la pression académique

Les niveaux de pression sont concentrés autour de **1, 3 et 5**, ce qui peut indiquer des groupes distincts :

- Certains étudiants ressentent **une pression faible (1)**.
- D'autres perçoivent **une pression modérée (3)**.
- Certains ont une pression **très forte (5)**, ce qui pourrait influencer leur santé mentale.

#### Distribution de la moyenne des notes

La majorité des notes sont comprises **entre 5 et 10**, indiquant une répartition équilibrée des performances académiques.

#### Distribution de la satisfaction aux études

La satisfaction est principalement **groupée autour de 1, 3 et 5**, montrant des disparités nettes :

- Certains étudiants sont **très satisfaits (5)**.
- D'autres sont **peu satisfaits (1)**, ce qui pourrait être lié à leur niveau de pression académique.

### Distribution du nombre d'heures de travail ou d'étude

La répartition semble **relativement équilibrée**, ce qui suggère que les étudiants consacrent des efforts similaires à leurs études.

### Distribution du stress financier

Les niveaux de stress sont **groupés autour de 1, 3 et 5**, ce qui indique des perceptions très distinctes :

- Certains étudiants ressentent **peu de stress financier (1)**.
- D'autres subissent une **pression modérée (3)** ou **élevée (5)**, ce qui pourrait impacter leur santé mentale.

## 2) Analyse univariée des variables catégorielles

### 2.1) Proportions et Interprétations

Variable	Modalités principales	Proportions principales	Interprétation
sex	Male : 15 546 Female : 12 352	56 % 44 %	La population étudiée est légèrement majoritairement masculine.
durée de sommeil	<5h : 8 309 7–8h : 7 346 5–6h : 6 181 >8h : 6 044	30 % 26 % 22 % 22 %	Près d'1 étudiant sur 3 dort <5h ; peu dorment suffisamment.
habitudes alimentaires	Unhealthy : 10 316 Moderate : 9 921 Healthy : 7 649	37 % 36 %	Plus de 70 % ont une alimentation peu équilibrée.
pensées suicidaires	Yes : 17 656 No : 10 242	63 % 37 %	Plus de 60 % des étudiants ont déjà eu des pensées suicidaires.
antécédents familiaux (maladie mentale)	No : 14 397 Yes : 13 501	52 % 48 %	Une part importante d'étudiants ont des antécédents familiaux.
dépression	Yes : 16 335 No : 11 563	59 % 41 %	Près de 6 étudiants sur 10 sont en dépression, ce qui est alarmant.

Tableau 8 : Tableau des proportions des variables catégorielles

## 2.2) Représentations graphiques

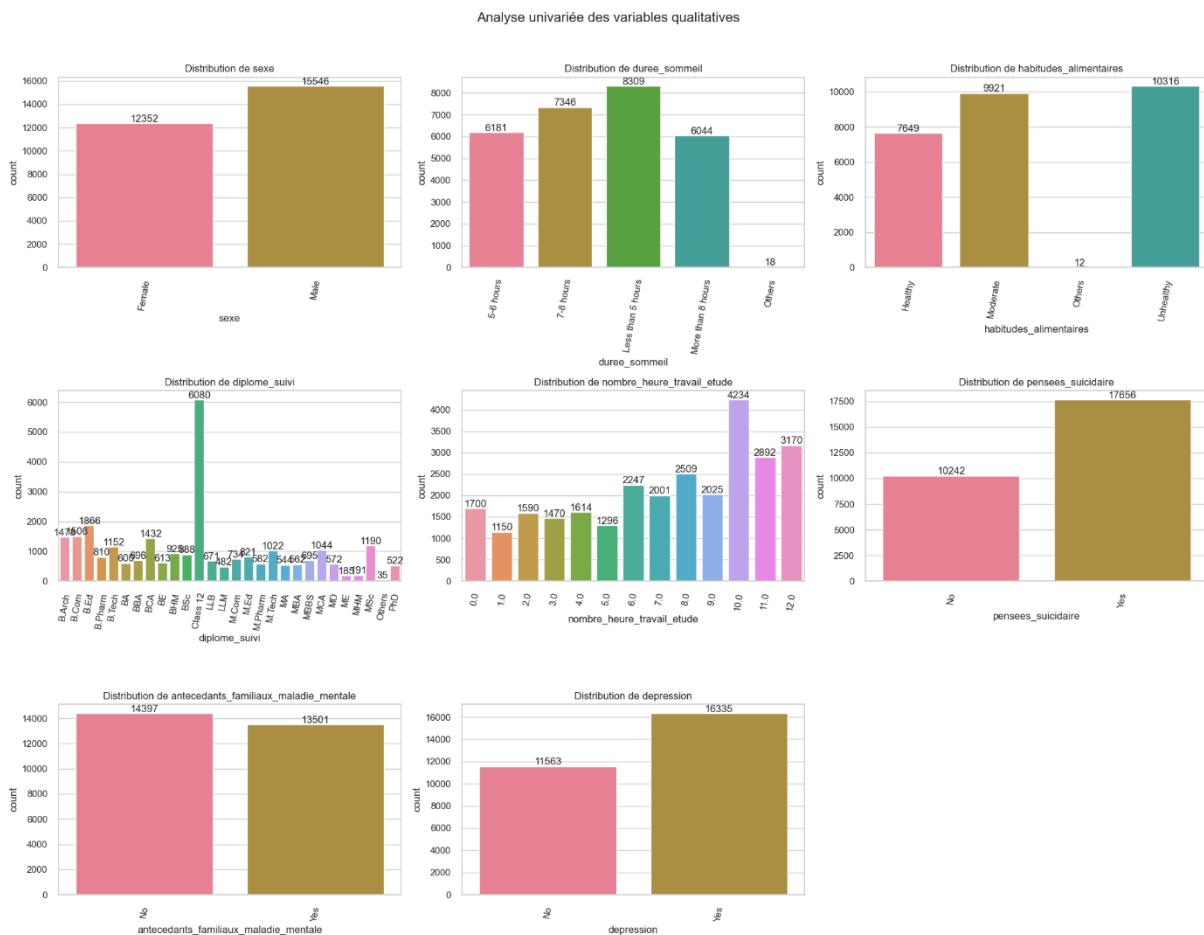


Figure 32: Représentations graphique des variables catégorielles

### Interprétation du graphique

Ce graphique représente plusieurs distributions essentielles des données étudiées :

- Sexe** : Une légère prédominance des étudiants masculins par rapport aux féminins.
- Durée du sommeil** : La majorité des étudiants dorment entre 5 et 8 heures, avec peu de cas de sommeil très long ou très court.
- Habitudes alimentaires** : Une répartition équilibrée entre les habitudes saines, modérées et malsaines, suggérant des différences potentielles dans l'impact sur la santé mentale.
- Diplômes suivis** : Les diplômes les plus représentés regroupent les étudiants inscrits en Class 12 et B.Ed, ce qui suggère une majorité d'étudiants dans ces filières.
- Nombre d'heures de travail ou d'étude** : La distribution montre une variation importante, avec un pic autour de 10 heures par jour.
- Pensées suicidaires** : Une proportion significative d'étudiants déclare en avoir eu, renforçant l'importance de l'analyse de la santé mentale.
- Antécédents familiaux de maladies mentales** : Répartition quasi équitable entre ceux ayant des antécédents et ceux qui n'en ont pas.

- **Dépression** : Une part importante d'étudiants est concernée, soulignant l'impact des divers facteurs étudiés.

## Visualisation spéciale pour les variables clés

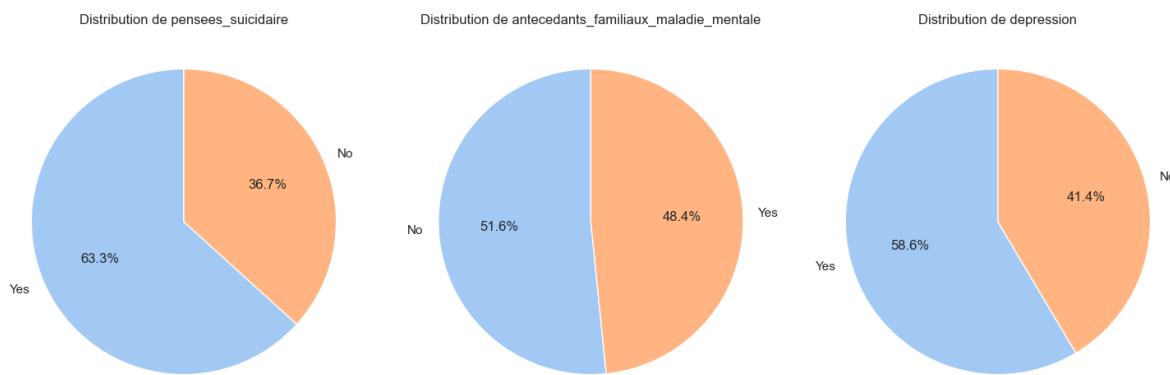


Figure 33: Répartition des variables catégorielles clés

### 📌 Interprétation des Distributions

#### 1. Distribution des Pensées Suicidaires

La répartition des individus selon les pensées suicidaires révèle que :

- 63.3% n'ont jamais eu de pensées suicidaires (No),
- 36.7% ont déclaré avoir eu des pensées suicidaires (Yes).

#### Interprétation :

Près de **4 individus sur 10** ont été confrontés à des pensées suicidaires, ce qui souligne l'importance critique de cette problématique dans la population étudiée. Cette proportion élevée justifie des interventions ciblées en prévention et en soutien psychologique.

#### 2. Distribution des Antécédents Familiaux de Maladie Mentale

Les données montrent une répartition quasi équilibrée :

- 51.6% sans antécédents familiaux (No),
- 48.4% avec antécédents familiaux (Yes).

#### Interprétation :

La présence d'antécédents familiaux chez près de la moitié des individus (48.4%) en fait un **facteur de risque potentiel** à investiguer. Une analyse croisée avec d'autres variables (ex : dépression) permettrait d'évaluer son impact réel.

#### 3. Distribution de la Dépression

La prévalence de la dépression se répartit comme suit :

- **58.6%** des individus sont non dépressifs (*No*),
- **41.4%** présentent des symptômes dépressifs (*Yes*).

#### Interprétation :

Avec **plus de 40%** de la population touchée, la dépression apparaît comme un enjeu majeur de santé publique dans cet échantillon. Cette proportion suggère un besoin accru de ressources dédiées au diagnostic et à la prise en charge.



Figure 34: Boxplots des variables numériques clés

#### 💡 Interprétation des distributions (Boxplots)

Les trois boxplots ci-dessus représentent la distribution des variables moyenne des notes, pression académique et stress financier chez les étudiants.

- **Moyenne des notes** : La distribution est centrée autour de **7,8**, avec un intervalle interquartile allant de **6,3 à 8,9**. Aucun outlier n'est apparent. Cela indique que la majorité des étudiants ont des **résultats académiques globalement bons**, avec une concentration des notes dans la moyenne supérieure.
- **Pression académique** : Le boxplot montre une distribution **relativement symétrique**, avec une médiane située à **3** sur une échelle de 1 à 5. La majorité des réponses se situent entre **2 et 4**, ce qui montre que les étudiants ressentent **une pression modérée**, sans extrêmes notables.
- **Stress financier** : La médiane est également à **3**, ce qui suggère un **stress financier modéré** pour une large partie des étudiants. L'intervalle interquartile est situé entre **2 et 4**, ce qui reflète une certaine diversité des situations économiques sans présence apparente de valeurs aberrantes.

### III) ANALYSE BIVARIEE

#### Normalité

##### ❖ Test graphique

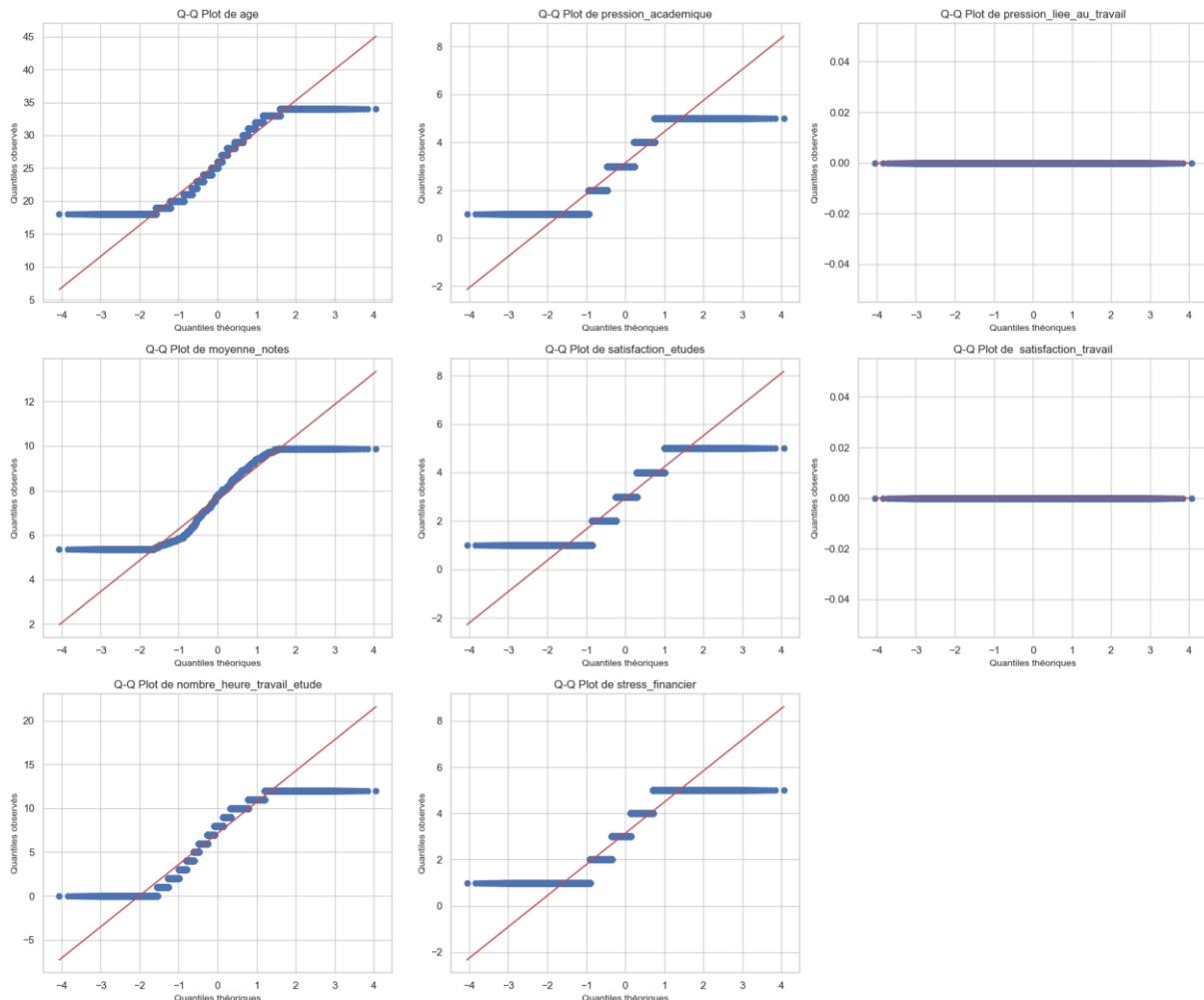


Figure 35: Test graphique de normalité

Ce graphique présente une série de **Q-Q plots (Quantile-Quantile plots)** qui comparent la distribution de différentes variables à une distribution normale.

Nous constatons qu'aucune variable ne suit une distribution normale excepté les variables pression\_liée\_au\_travail et satisfaction\_travail où tous les points sont alignés horizontalement, indiquant probablement une **valeur constante**. Ces variables semblent **constantes** (pas de variance), **non exploitable pour des tests paramétriques**.

## ❖ Test statistique

Variable	Test de Shapiro-Wilk (p-val)	Test KS (p-val)	Test AD (stat)	Test JB (p-val)	Conclusion globale
age	0.0000	0.0000	368.0620	0.0000	Normalité rejetée
pression_academique	0.0000	0.0000	970.6121	0.0000	Normalité rejetée
pression_liee_au_travail	1.0000	0.0000	NaN	NaN	Normalité acceptée (const.)
moyenne_notes	0.0000	0.0000	435.5696	0.0000	Normalité rejetée
satisfaction_etudes	0.0000	0.0000	943.6700	0.0000	Normalité rejetée
satisfaction_travail	1.0000	0.0000	NaN	NaN	Normalité acceptée (const.)
nombre_heure_travail_etude	0.0000	0.0000	658.2001	0.0000	Normalité rejetée
stress_financier	0.0000	0.0000	1084.0890	0.0000	Normalité rejetée

Tableau 9: Test statistique de normalité

## 1) Pour une variable quantitative et qualitative

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

L'âge ne suit pas la loi normale, la variable dépression étant binaire, nous allons utiliser le test non paramétrique de Mann Whitney U adapté à cette analyse, dont les hypothèses sont :

H0 : Pas de liaison significative

H0 : Liaison significative

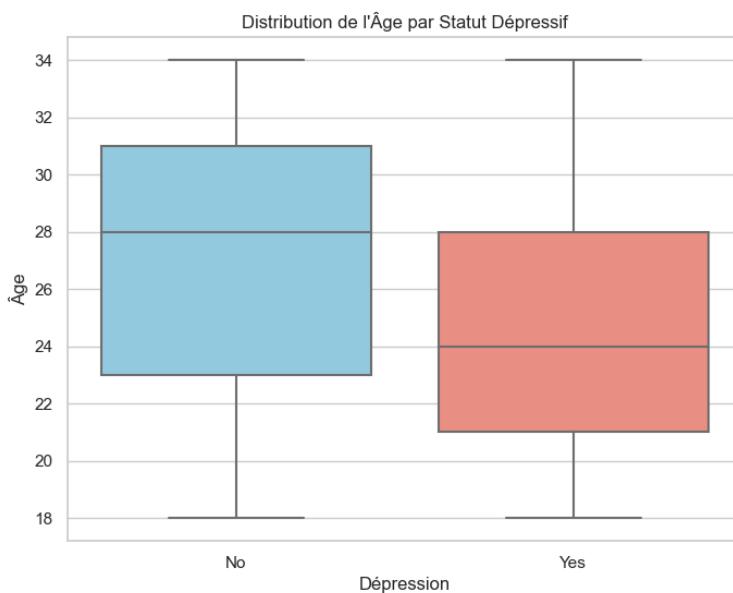


Figure 36: Boxplot de l'âge par statut dépressifs

Cette visualisation suggère que, dans cet échantillon, les personnes souffrant de dépression tendent à être un peu plus jeunes en moyenne comparées aux personnes non dépressives. La répartition des âges est également moins étendue dans le groupe dépressif. Ces différences montrent que les jeunes sont plus exposés à certains facteurs de risque de dépression de cet échantillon.

Le **Test de Mann Whitney U** nous donne une p-value = 0.0000 indiquant une liaison significative entre l'âge et la dépression. Suggérant une influence marquée de l'âge sur la dépression.

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

Le graphique illustre la relation entre la pression académique et le statut dépressif des étudiants. Il montre que ceux qui déclarent être en état de dépression ressentent une pression académique plus élevée en moyenne par rapport à ceux qui ne sont pas dépressifs. La médiane de la pression académique est plus haute pour les étudiants en situation de dépression. Cette tendance suggère une association entre le stress académique et la présence de troubles dépressifs.

Le **Test de Mann Whitney U** nous donne une p-value = 0.0000 indiquant une liaison significative entre la pression académique et la dépression. Suggérant une influence considérable de celle-ci sur la dépression.



Figure 37: Boxplot pression\_académique par dépressifs

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

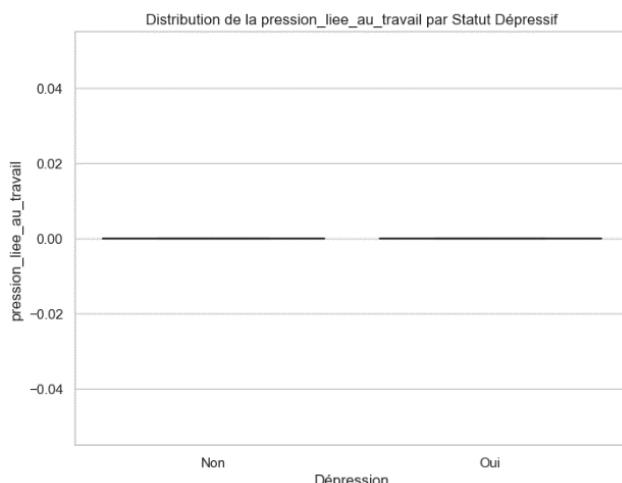


Figure 38: Boxplot pression\_liée\_travail par dépressifs

- **Analyse de la relation entre satisfaction travail et la dépression**



Figure 39: Boxplot de satisfaction travail par dépressifs

Le graphique examine la relation entre la pression liée au travail et le statut dépressif. Il indique que la pression liée au travail reste à zéro, indépendamment du statut dépressif des étudiants. Autrement dit, il n'y a pas de différence significative entre les personnes déclarant être en dépression et celles qui ne le sont pas en termes de pression de travail ressentie. Ce constat suggère que la dépression ne semble pas être directement influencée par la pression au travail dans cet ensemble de données.

Le graphique examine la relation entre la satisfaction travail et le statut dépressif. Il indique que la satisfaction travail reste à zéro, indépendamment du statut dépressif des étudiants. Autrement dit, il n'y a pas de différence significative entre les personnes déclarant être en dépression et celles qui ne le sont pas en termes de satisfaction travail ressentie. Ce constat suggère que la dépression ne semble pas être directement influencée par la pression au travail dans cet ensemble de données.

- Analyse de la relation entre la moyenne des notes et la dépression

Ce graphique représente la distribution de la moyenne des notes en fonction du statut dépressif. On observe que les individus ne souffrant pas de dépression ont une médiane des notes légèrement supérieure à celle des individus dépressifs. Cependant, la dispersion des notes est similaire entre les deux groupes.

Le **Test de Mann Whitney U** nous donne une p-value = 0.0003 indiquant une liaison significative entre la moyenne des notes et la dépression. Suggérant une grande influence des notes sur la dépression.

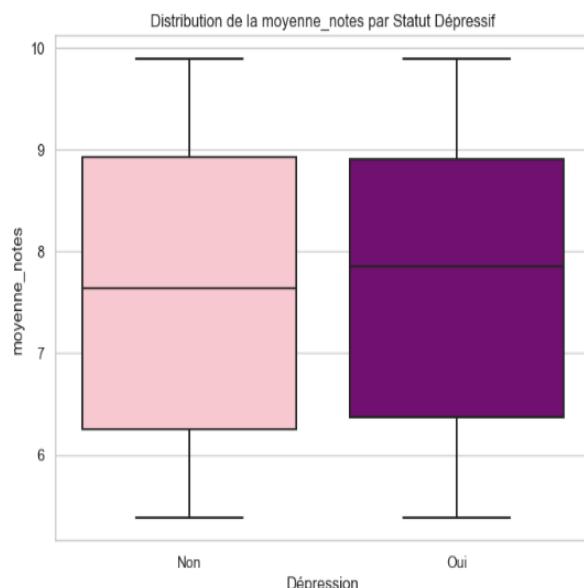


Figure 40: Boxplot moyenne\_notes par dépressifs

- Analyse de la relation entre satisfaction étude et la dépression

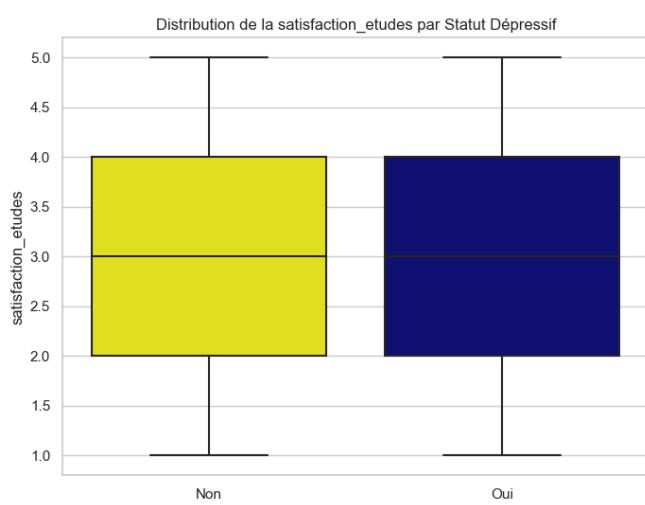


Figure 41: Boxplot satisfaction étude par dépressifs

Le graphique illustre la distribution de la satisfaction des études en fonction du statut dépressif. Les médianes sont de 3,0 pour les deux groupes, et la dispersion des valeurs est similaire, avec des quartiles allant de 2,0 à 4,0 et des moustaches s'étendant de 1,0 à 5,0.

Cette représentation montre que la satisfaction académique ne semble pas être significativement influencée par le statut dépressif, puisque les tendances restent les mêmes entre les groupes.

Le **Test de Mann Whitney U** montre une différence significative entre la satisfaction aux études et la dépression avec une p-value = 0.0000. Cela suggère que la satisfaction aux études est liée à la dépression des étudiants.

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

Le graphique illustre la distribution du nombre d'heures consacrées au travail et aux études en fonction du statut dépressif. Il met en évidence une tendance : les étudiants en situation de dépression déclarent en moyenne un volume d'heures de travail ou d'étude plus élevé que ceux qui ne sont pas dépressifs. La médiane pour le groupe dépressif est autour de 8 heures, contre environ 6 heures pour les non-dépressifs.

Cette observation suggère une possible corrélation entre une charge de travail plus importante et le statut dépressif

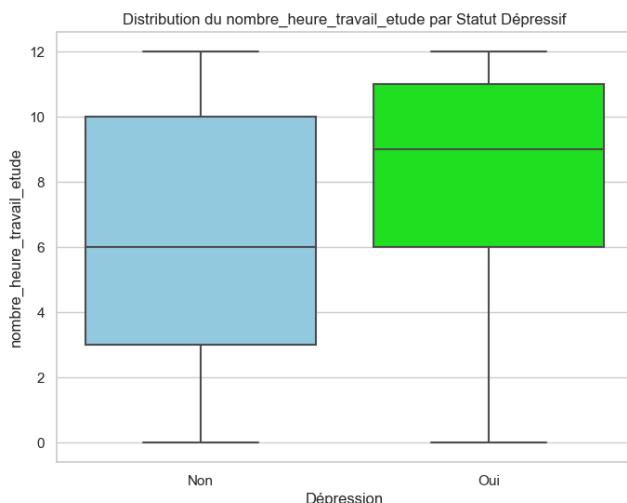


Figure 42: Boxplot du nombre\_heure\_travail\_etude par dépressifs

Le test de Mann-Whitney U indique une différence significative entre le nombre d'heures de travail ou d'étude et la dépression avec une p-value = 0.0000. Cela signifie que le temps consacré au travail ou aux études influencent la dépression chez les étudiants.

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

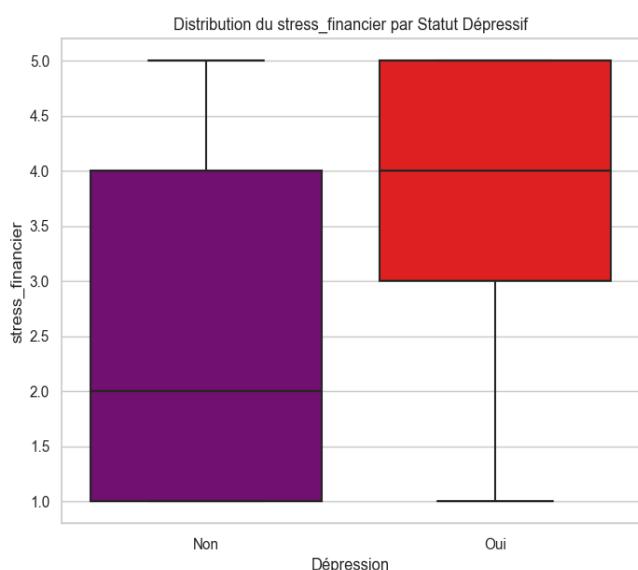


Figure 43: Boxplot du stress financier par dépressifs

Le test de Mann-Whitney U montre une liaison significative entre le stress financier et la dépression avec une p-value = 0.0000. Cela signifie que le niveau de stress financier est lié à la présence ou non de symptômes dépressifs chez les étudiants.

Le graphique met en évidence la distribution du stress financier selon le statut dépressif des individus. Il révèle une nette différence entre les deux groupes : les personnes en dépression déclarent un niveau de stress financier plus élevé, avec une médiane proche de 4, alors que celles sans dépression affichent une médiane autour de 2. La dispersion des valeurs montre également que le stress financier est plus variable chez les individus dépressifs, avec une tendance marquée vers des niveaux élevés.

Ce résultat suggère une association entre le stress financier et la présence de symptômes dépressifs, soulignant l'impact potentiel des contraintes économiques sur la santé mentale.

## 2) Pour deux variables qualitatives

### Condition de Cochran

Variable	% Cellules valides	Condition de Cochran respectée
pensees_suicidaire	100.0 %	Oui
habitudes_alimentaires	87.5 %	Oui
diplome_suivi	100.0 %	Oui
duree_sommeil	100.0 %	Oui
antecedants_familiaux_maladie_mentale	100.0 %	Oui
ville	57.7 %	✗ Non
profession	7.1 %	✗ Non
sexe	100.0 %	Oui

Tableau 10: Table de vérification de la condition de Cochran

- Analyse de la relation entre le sexe et la dépression

Dépression	Femme	Homme	Total
Non	5 132	6 431	11 563
Oui	7 220	9 115	16 335
Total	12 352	15 546	27 898

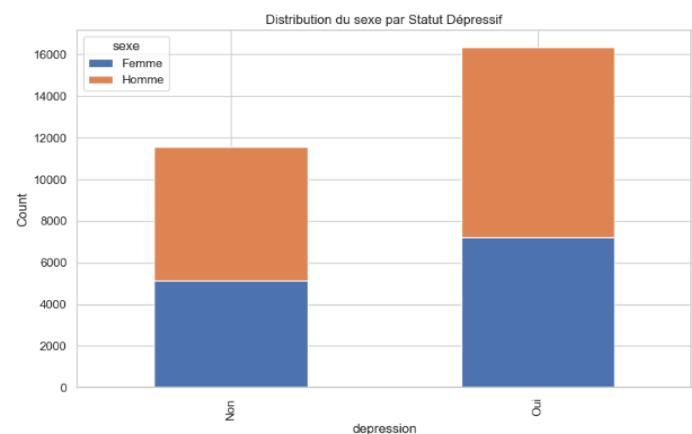


Figure 44: Distribution du sexe par dépressifs

La variable sexe respecte la condition de Cochran, alors nous allons utiliser le **test de Chi<sup>2</sup>** pour analyser la relation entre le sexe et la dépression. Les hypothèses de ce test sont :

H<sub>0</sub> : les deux variables ne sont pas liées

H<sub>1</sub> : les deux variables sont liées

Le test du Chi<sup>2</sup> entre le sexe et la dépression donne une **p-value de 0.771**. Comme cette valeur est largement supérieure à 0.05, on ne rejette pas l'hypothèse nulle. **Cela signifie qu'il n'y a pas de lien significatif entre le sexe et la présence de symptômes dépressifs chez les étudiants.**

### • Analyse de la relation entre la ville et la dépression

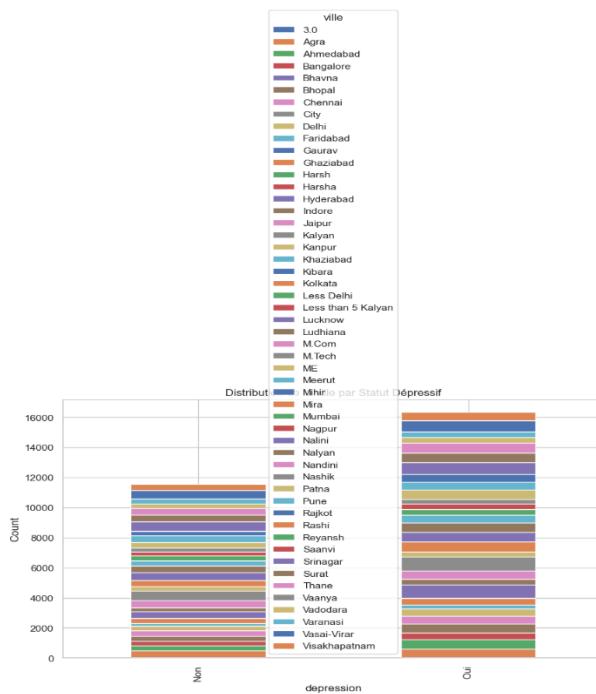
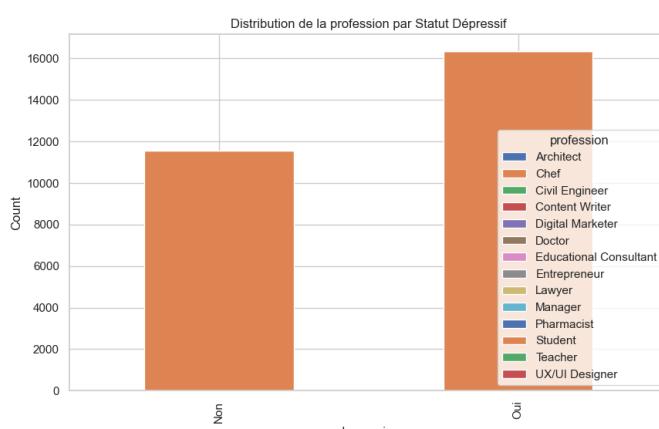


Figure 45: Distribution de la ville par dépressifs

Ville	Non dépressif	Dépressif
Agra	509	585
Ahmedabad	311	640
Bangalore	300	467
Bhavnagar	0	2
Bhopal	355	579
Chennai	357	528
City	1	1
Delhi	300	468
Faridabad	190	271
Reyansh	1	0
Saanvi	0	2
Srinagar	609	763
Surat	453	625
Thane	466	673
Vaanya	0	1
Vadodara	304	390
Varanasi	321	363

Le test du Chi<sup>2</sup> entre la ville et la dépression donne une **p-value de 0.000**, ce qui est bien en dessous du seuil de 0.05. Cela signifie qu'on rejette l'hypothèse selon laquelle il n'y aurait pas de lien entre la ville de résidence et la dépression. Autrement dit, la fréquence de la dépression varie de façon significative selon les villes. Ce résultat suggère que certains environnements urbains pourraient être associés à un niveau plus élevé ou plus faible de symptômes dépressifs.

### • Analyse de la relation entre la profession et la dépression



Profession	% Dépressifs	% Non dépressifs
Architect	87.5 %	12.5 %
Chef	100.0 %	0.0 %
Civil Engineer	100.0 %	0.0 %
Content Writer	100.0 %	0.0 %
Digital Marketer	66.7 %	33.3 %
Doctor	100.0 %	0.0 %
Educational Consultant	100.0 %	0.0 %
Entrepreneur	100.0 %	0.0 %
Lawyer	100.0 %	0.0 %
Manager	100.0 %	0.0 %
Pharmacist	100.0 %	0.0 %
Student	58.5 %	41.5 %
Teacher	83.3 %	16.7 %
UX/UI Designer	100.0 %	0.0 %

Figure 46: Distribution de la profession et du statut dépressif

Le test du Chi<sup>2</sup> entre la profession et la dépression donne une **p-value de 0.354**. Cela signifie qu'il n'y a pas de lien statistiquement significatif entre ces deux variables. Autrement dit, **le type de profession exercée n'a pas avoir d'impact** direct sur la présence de dépression chez les participants.

- Analyse de la relation entre la durée de sommeil et la dépression

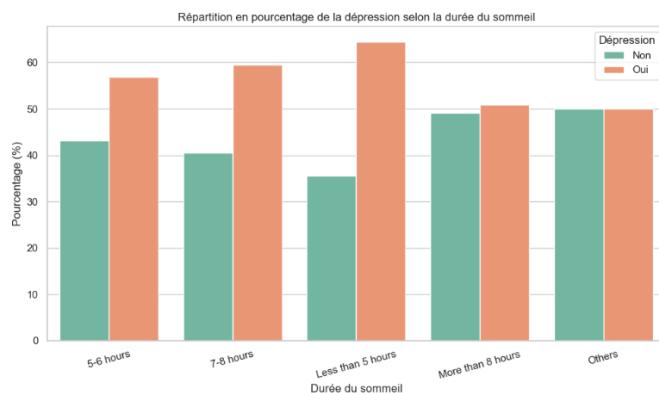


Figure 47: Distribution de la durée de sommeil par dépressifs

Le test du Chi2 entre la durée de sommeil et la dépression donne une **p-value de 0.000**. Cela signifie qu'il existe une relation statistiquement significative entre ces deux variables. En d'autres termes, **la durée de sommeil est liée à l'état de dépression** : certaines durées de sommeil sont plus fréquentes chez les personnes dépressives que chez les non-dépressives.

- Analyse de la relation entre les habitudes alimentaires et la dépression

Habitudes alimentaires	% Dépressifs	% Non dépressifs
Healthy	45.4?	54.6?
Moderate	56.0?	44.0?
Unhealthy	70.7?	29.3?
Others	66.7?	33.3?

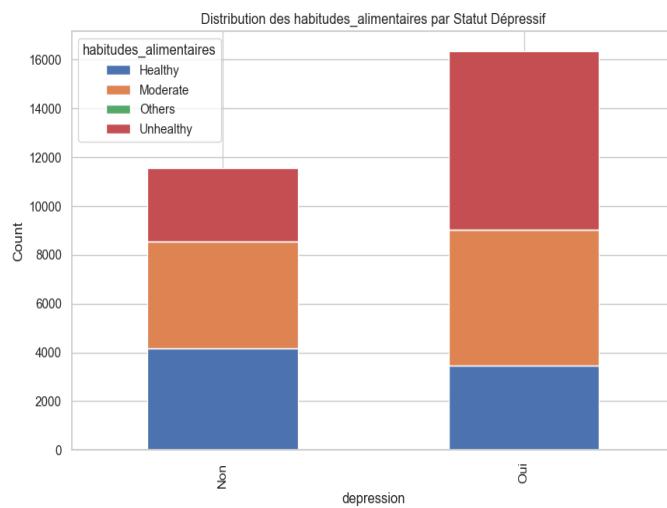


Figure 48: Distribution des habitudes alimentaires par dépressifs

Le test du Chi2 entre les habitudes alimentaires et la dépression a donné une **p-value de 0.000**, ce qui indique une relation statistiquement significative entre ces deux variables. Autrement dit, **les habitudes alimentaires sont liées à l'état de dépression**. Par exemple, parmi les personnes ayant une alimentation "unhealthy", 70.7 % sont dépressives, contre seulement 45.4 % chez celles ayant une alimentation "healthy". Ce résultat suggère que de mauvaises habitudes alimentaires pourraient être associées à un risque plus élevé de dépression.

- Analyse de la relation entre le diplôme suivi et la dépression

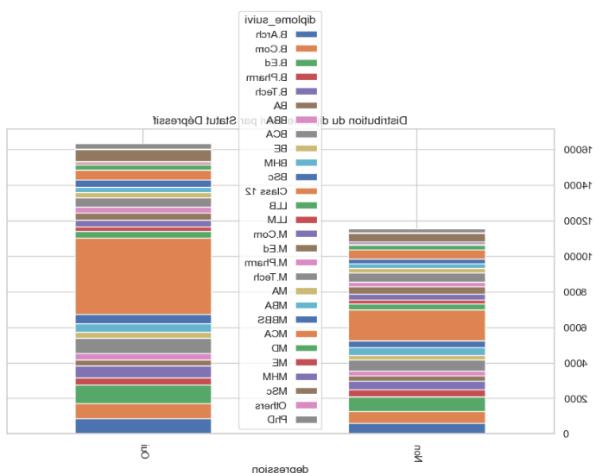


Figure 49: Distribution du diplôme suivi par dépressifs

Le test du Chi<sup>2</sup> entre le diplôme suivi et la dépression a donné une **p-value de 0.000**, ce qui indique une **relation statistiquement significative** entre ces deux variables. Cela signifie que le taux de dépression varie selon le type de diplôme. Par exemple, 60 % des individus classés dans la catégorie "Others" sont dépressifs, contre environ 52 % chez ceux ayant un diplôme de type "MD" ou "MHM". Ce résultat suggère que le niveau ou le type d'étude suivi peut être lié au risque de dépression.

Diplôme suivi	% Dépressifs	% Non dépressifs
B.Arch	58.9 %	41.1 %
B.Com	56.6 %	43.4 %
B.Ed	54.7 %	45.3 %
B.Pharm	52.8 %	47.2 %
B.Tech	56.8 %	43.2 %
BA	53.5 %	46.5 %
BBA	58.5 %	41.5 %
BCA	57.1 %	42.9 %
BE	54.5 %	45.5 %
BHM	55.0 %	45.0 %
MA	53.3 %	46.7 %
MBA	53.9 %	46.1 %
MBBS	58.1 %	41.9 %
MCA	53.5 %	46.5 %
MD	52.1 %	47.9 %
ME	53.0 %	47.0 %
MHM	51.8 %	48.2 %
MSc	57.0 %	43.0 %
Others	60.0 %	40.0 %
DPL	54.0 %	45.0 %

- Analyse de la relation entre la pensées suicidaire et la dépression

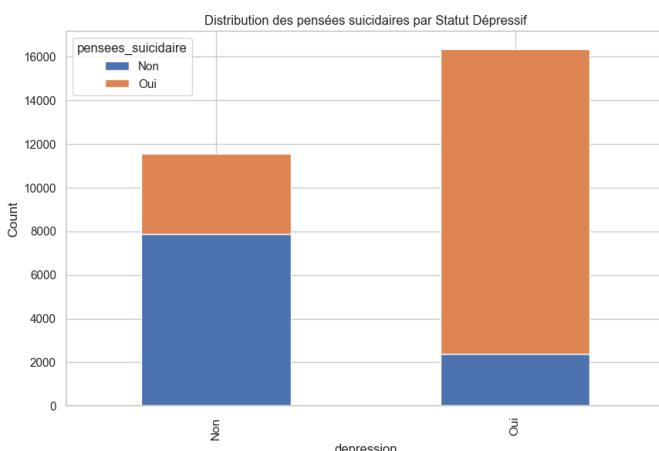


Figure 50: Distribution de la pensée suicidaire par dépressifs

Pensées suicidaires	% Non dépressifs	% Dépressifs
Non	68.04 %	14.56 %
Oui	31.96 %	85.44 %

Le test du Chi<sup>2</sup> mené entre les pensées suicidaires et la dépression a donné une **p-value égale à 0.000**, ce qui est inférieure au seuil de signification de 0.05. Cela signifie que nous rejetons l'hypothèse nulle ( $H_0$ ), selon laquelle il n'existe **aucun lien** entre ces deux variables.

Nous concluons donc qu'il existe **une liaison statistiquement significative** entre les pensées suicidaires et la dépression. En d'autres termes, les étudiants ayant des pensées suicidaires sont plus susceptibles de faire une dépression, ce qui met en évidence l'importance de surveiller cet indicateur dans les stratégies de prévention.

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

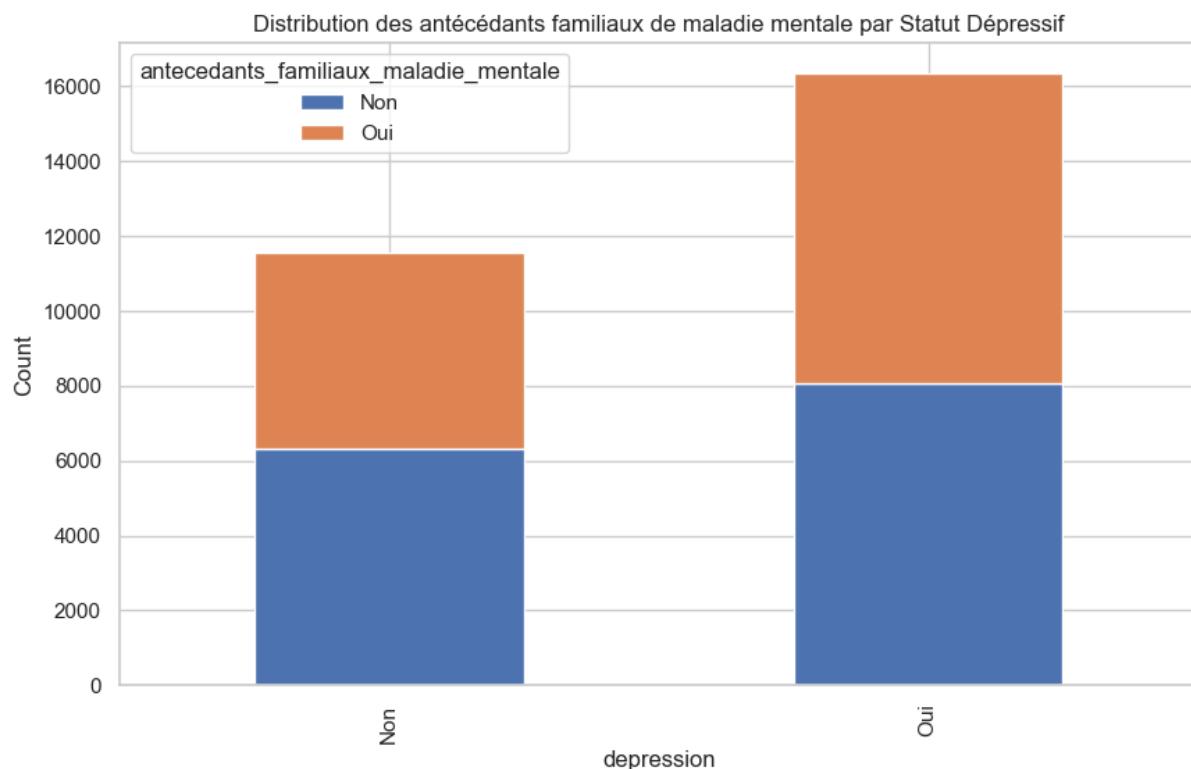


Figure 51: Distribution des antécédents familiaux de maladie mentale par dépressifs

Antécédents familiaux	% Non dépressifs	% Dépressifs
Non	54.74 %	49.37 %
Oui	45.26 %	50.63 %

Le test du Chi<sup>2</sup> réalisé entre les antécédents familiaux de maladie mentale et la dépression a donné une **p-value de 0.000**, ce qui est bien en dessous du seuil de 0.05. Cela signifie que nous rejetons l'hypothèse nulle ( $H_0$ ) et que les deux variables sont **statistiquement liées**. En d'autres termes, il existe une **relation significative** entre le fait d'avoir des antécédents familiaux de troubles mentaux et la présence de dépression. Les étudiants ayant de tels antécédents sont donc plus à risque de souffrir de dépression, ce qui souligne l'importance de prendre en compte l'historique familial dans l'évaluation psychologique.

## **IV) STATISTIQUE INFÉRENTELLE**

### **1) Trouvons l'intervalle de confiance pour la proportion d'étudiants ayant déjà eu des pensées suicidaires**

Pour résoudre ce problème, la méthode de **bootstrap** a été utilisée pour estimer l'intervalle de confiance de la proportion d'étudiants ayant déjà eu des pensées suicidaires. Bien que cette variable soit binaire et ne nécessite pas de condition de normalité, le bootstrap reste pertinent car le bootstrap ne suppose aucune hypothèse sur la forme de la distribution de la proportion. Il se base uniquement sur les résultats observés dans l'échantillon, donc il reste précis et robuste, même si la proportion est très faible ou très élevée. Cette approche empirique permet ainsi d'obtenir un intervalle de confiance plus fiable et réaliste que les méthodes classiques basées sur l'approximation normale.

Ainsi, avec un niveau de confiance de 95 %, la vraie proportion d'étudiants ayant déjà eu des pensées suicidaires se situe entre **62.72%** et **63.84%**, ce qui suggère une problématique préoccupante à considérer dans l'analyse des facteurs de dépression.

### **2) Estimation de la moyenne et la médiane des heures de travail ou d'études pour les étudiants souffrant de dépression.**

Chez les étudiants qui déclarent un état dépressif, le nombre moyen d'heures consacrées au travail ou aux études est estimé, par la méthode du bootstrap, à **7,8 heures dans l'intervalle [7,8, 7,9]** avec un niveau de confiance de 95 %. Par ailleurs, la médiane est estimée au seuil de 95% à **9 heures**, avec un intervalle de confiance très resserré [9,0, 9,0], ce qui indique que la moitié des étudiants dépressifs travaillent ou étudient **9 heures par jour**. Ce résultat suggère une charge de travail élevée, qui pourrait être un facteur lié à leur état psychologique.

### **3) Évaluons la moyenne et la médiane du stress financier pour les étudiants avec et sans dépression.**

Les résultats indiquent une différence marquée de stress financier entre les étudiants souffrant de dépression et ceux qui n'en présentent pas. Chez les étudiants dépressifs, la moyenne et la médiane du stress financier sont respectivement de **3,6** et **4,0**, avec des intervalles de confiance à **95 %** très resserrés : [3,6 ; 3,6] pour la moyenne et [4,0 ; 4,0] pour la médiane. En comparaison, les étudiants non dépressifs présentent une moyenne de **2,5** et une médiane de **2,0**, avec des intervalles à **95 %** également stables : [2,5 ; 2,5] et [2,0 ; 2,0]. Ces résultats suggèrent que le stress financier est significativement plus élevé chez les étudiants dépressifs, ce qui en fait un facteur potentiellement associé à leur état psychologique.

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

Le test de Mann-Whitney U a été réalisé afin d'évaluer si la satisfaction des études diffère significativement entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas. La p-value obtenue est inférieure à 0.0001, ce qui est largement en dessous du seuil de signification de 5 %. Cela conduit au **rejet de l'hypothèse nulle ( $H_0$ )** selon laquelle les deux groupes auraient des niveaux de satisfaction similaires.

Ainsi, on conclut qu'il existe une **différence statistiquement significative** de satisfaction vis-à-vis des études entre les deux groupes, suggérant que la dépression pourrait être associée à un niveau de satisfaction académique plus faible.

#### 5) Les niveaux de satisfaction au travail diffèrent-ils significativement selon le diplôme suivi ?

**Les niveaux de satisfaction au travail ne diffèrent pas selon le diplôme suivi,** car toutes les valeurs observées pour cette variable sont identiques (égales à zéro) pour l'ensemble des étudiants, quel que soit leur niveau de diplôme. Cette absence totale de variabilité rend impossible toute comparaison statistique, y compris par le test de Kruskal-Wallis. Ainsi, il n'est pas possible de conclure à une différence de satisfaction au travail entre les groupes de diplômes suivis.

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

Le test du Chi<sup>2</sup> d'indépendance a été appliqué pour évaluer la relation entre la dépression et les habitudes alimentaires. Les deux variables étant qualitatives, ce test est approprié, et la **condition de validité de Cochran** a été vérifiée (toutes les fréquences attendues sont supérieures à 5), justifiant son utilisation.

Les résultats montrent une **statistique de Chi<sup>2</sup> de 1203,27** avec une **p-value < 0,0001**, ce qui est largement inférieur au seuil de signification de 5 %.

On rejette donc l'hypothèse nulle d'indépendance entre les deux variables. Autrement dit, il existe une **relation statistiquement significative** entre la dépression et les habitudes alimentaires. Cela suggère que la fréquence de dépression varie selon le type d'habitudes alimentaires adoptées par les étudiants.

### ❖ Carte de chaleur (Heatmap)

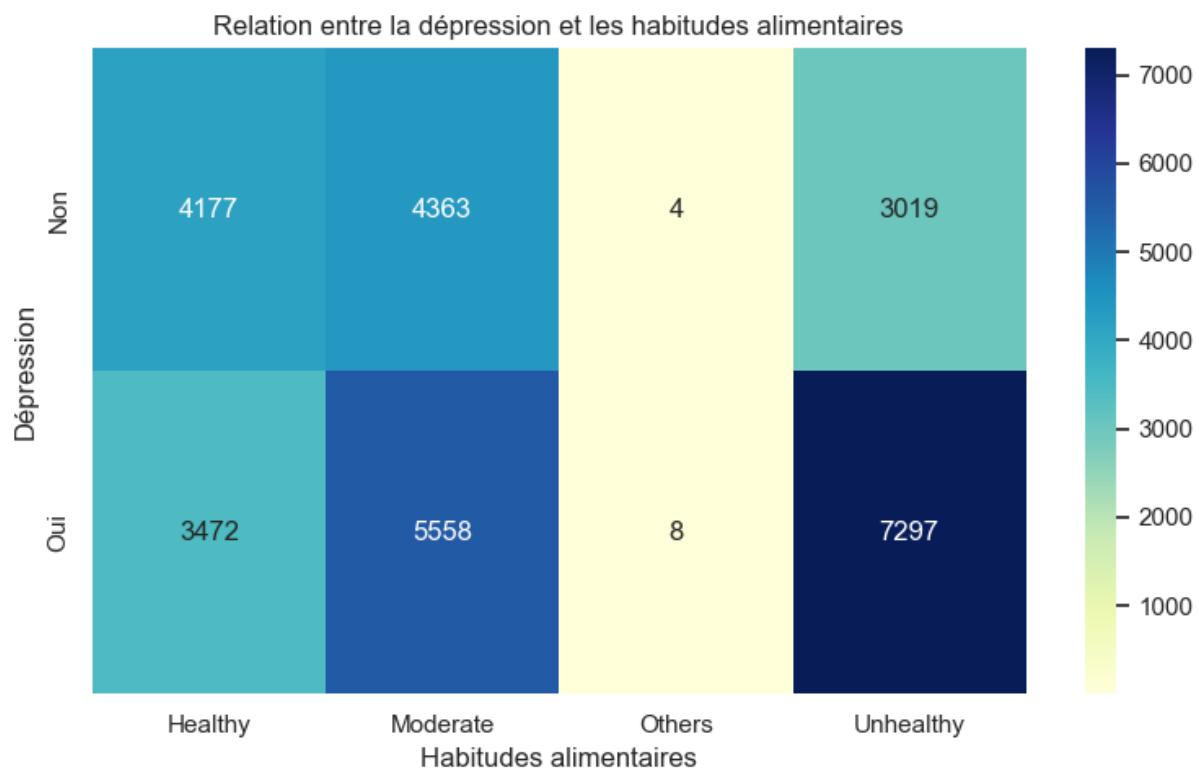


Figure 52: Répartition des cas de dépression par habitudes alimentaires

La heatmap ci-dessus illustre la répartition des cas de dépression en fonction des habitudes alimentaires. On observe une densité très marquée de cas de dépression chez les étudiants ayant des habitudes **Unhealthy (malsaines)** et **Moderate**, tandis que les étudiants aux habitudes **Healthy (saines)** présentent proportionnellement moins de cas de dépression. Ce résultat renforce la conclusion du test de Chi<sup>2</sup> : les habitudes alimentaires et la dépression sont liées de manière significative.

### 7) La durée du sommeil est-elle indépendante de la dépression ?

Le test du Chi<sup>2</sup> d'indépendance a été utilisé pour évaluer la relation entre la dépression et la durée du sommeil. La condition de validité du test (Cochran) a été vérifiée, et les résultats montrent une **statistique de Chi<sup>2</sup> de 277,14 avec une p-value inférieure à 0,0001**.

Cette p-value très faible conduit au **rejet de l'hypothèse nulle**, ce qui signifie qu'il existe une **relation statistiquement significative** entre la dépression et la durée du sommeil. Autrement dit, la fréquence de dépression varie en fonction du temps de sommeil déclaré par les étudiants.

Ce résultat suggère que la durée du sommeil est un facteur potentiellement associé à la dépression : des durées très courtes ou très longues pourraient être liées à une plus forte prévalence de symptômes dépressifs.

### ❖ Carte de chaleur (Heatmap)

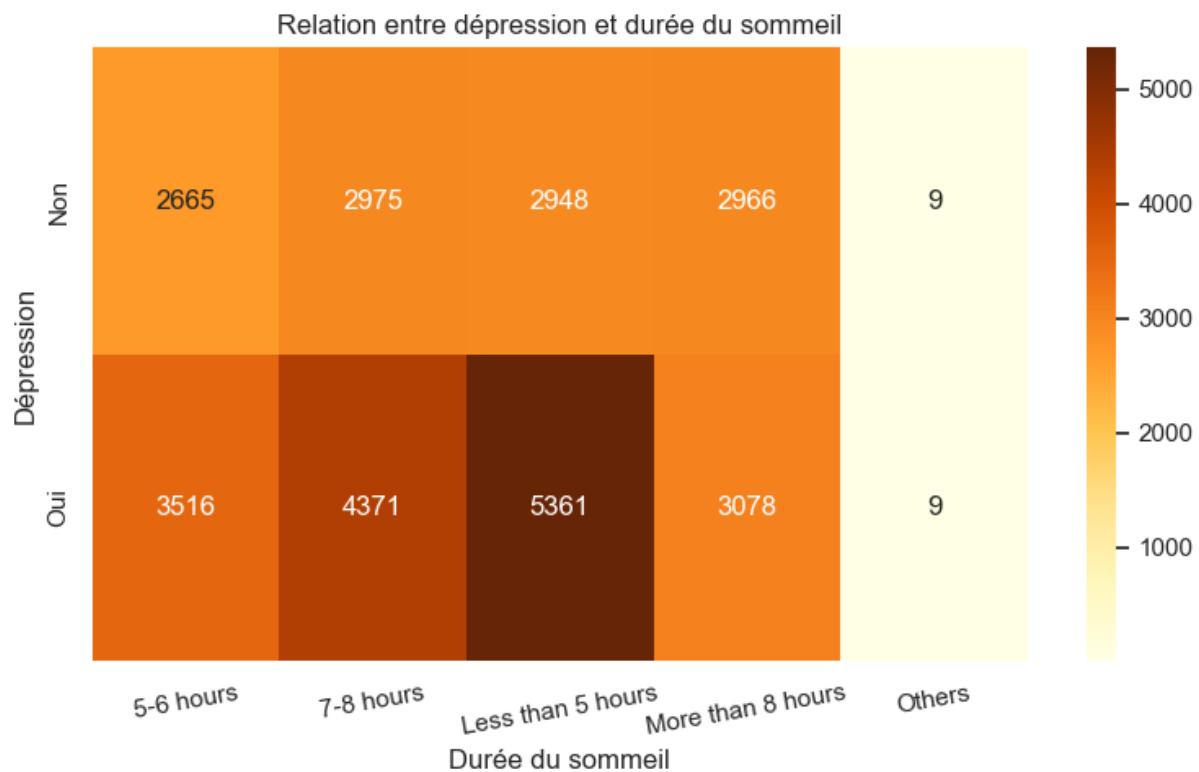


Figure 53: Répartition de la dépression en fonction de la durée de sommeil

La heatmap met en évidence une **distribution inégale de la dépression selon les différentes durées de sommeil**. On observe notamment que :

- Les étudiants dormant **moins de 5 heures** sont proportionnellement **beaucoup plus nombreux** à présenter des symptômes dépressifs (5361) comparé à ceux qui n'en souffrent pas (2948).
- Pour les durées de **7–8 heures**, qui correspondent souvent à une durée de sommeil recommandée, le nombre d'étudiants **sans dépression** (2975) est presque équivalent à celui **avec dépression** (4371), mais avec un léger déséquilibre en faveur du groupe dépressif.
- Les étudiants dormant **plus de 8 heures** semblent également touchés par la dépression (3078), bien que dans une moindre mesure.

Ces observations visuelles **renforcent les résultats du test du Chi<sup>2</sup>** : la durée du sommeil est significativement liée à la dépression. Des durées de sommeil **très courtes** (et possiblement très longues) sont associées à une **plus forte fréquence de symptômes dépressifs**.

## Conclusion

L'objectif principal de cette étude était d'analyser les facteurs associés à la dépression chez les étudiants, à travers une approche statistique inférentielle, afin de mieux comprendre les éléments susceptibles de contribuer à l'apparition ou à l'aggravation des symptômes dépressifs. Dans un contexte marqué par la montée des troubles psychologiques en milieu universitaire, cette analyse visait à éclairer les relations entre la dépression et plusieurs variables telles que les conditions académiques, les contraintes socio-économiques, les habitudes de vie ou encore le contexte personnel. En s'appuyant sur des tests non paramétriques et des méthodes d'estimation robustes, cette étude a permis de mettre en évidence des associations statistiquement significatives entre la dépression et plusieurs dimensions du quotidien étudiant.

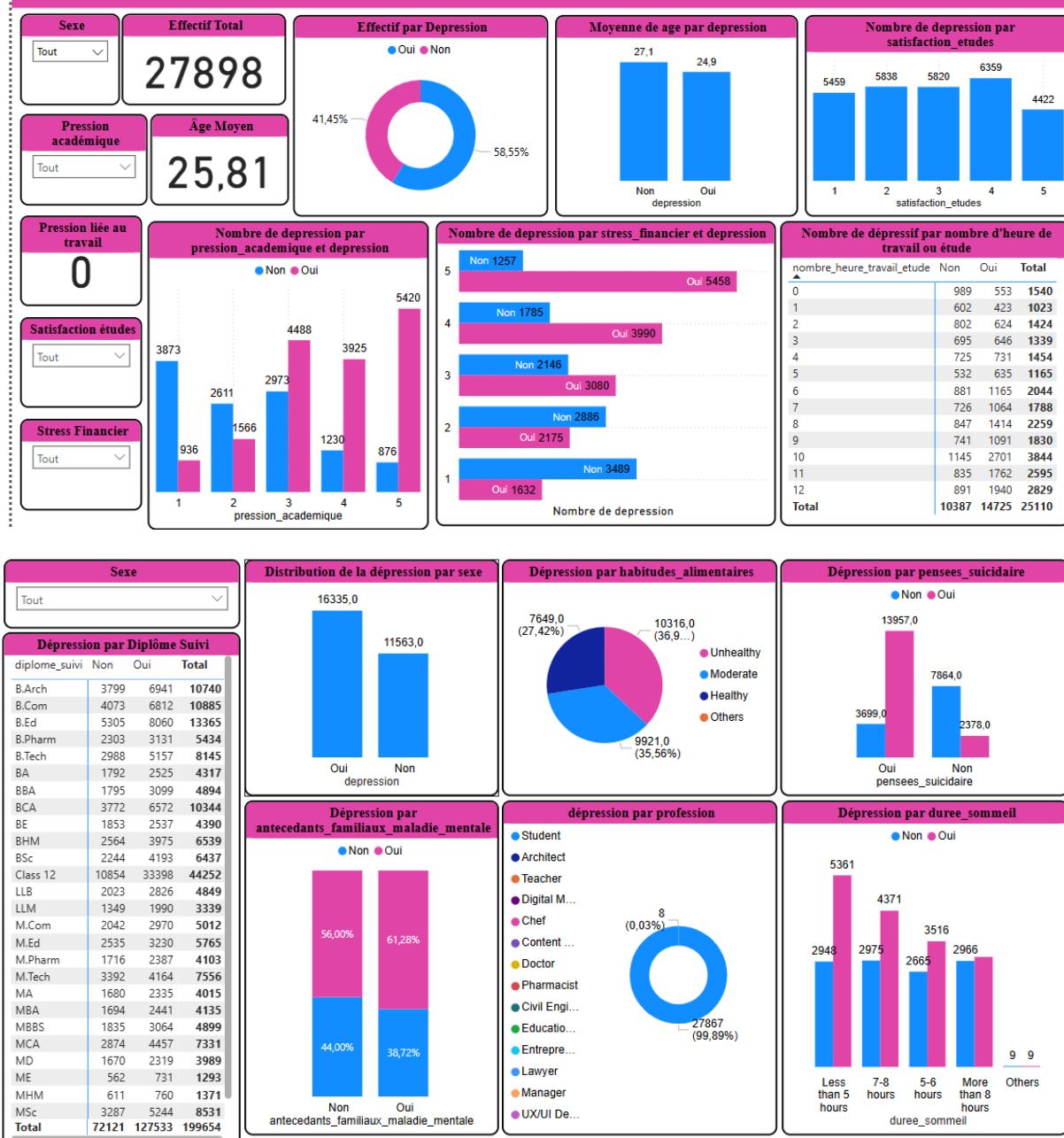
Les résultats ont révélé que certains facteurs, comme le stress financier, les habitudes alimentaires ou encore la durée du sommeil, sont fortement liés à la présence de symptômes dépressifs. De même, la satisfaction vis-à-vis des études ainsi qu'un temps de travail ou d'étude excessif apparaissent comme des éléments influents. L'analyse a également montré que les étudiants souffrant de dépression sont plus susceptibles d'avoir des habitudes de sommeil perturbées ou un stress financier élevé. À partir de ces constats, plusieurs recommandations peuvent être formulées : il est notamment essentiel de renforcer les dispositifs d'accompagnement psychologique dans les établissements, d'intégrer des campagnes de sensibilisation à l'hygiène de vie (sommeil, alimentation) et de mettre en place des mécanismes d'ajustement de la charge de travail pour les étudiants vulnérables. Une attention particulière devrait également être portée à la précarité financière des étudiants, qui constitue un facteur aggravant important.

Cependant, cette étude comporte certaines limites. La principale concerne la nature déclarative des données, qui peuvent être sujettes à des biais de réponse ou à une sous-estimation de certains symptômes. De plus, certaines variables ont présenté une absence totale de variabilité, comme la satisfaction au travail, ce qui a restreint la portée de certaines analyses. Par ailleurs, les tests utilisés ne permettent pas d'établir de relation causale, mais seulement des associations statistiques. Enfin, l'approche univariée ou bivariée utilisée ici gagnerait à être enrichie par des modèles multivariés intégrant des effets croisés.

Pour améliorer l'analyse, plusieurs pistes sont envisageables. L'utilisation de modèles de régression logistique pourrait permettre d'évaluer l'impact simultané de plusieurs facteurs sur la probabilité de développer une dépression. L'intégration de données qualitatives issues d'entretiens ou de questionnaires ouverts permettrait d'enrichir l'interprétation des résultats quantitatifs. Par ailleurs, une approche longitudinale pourrait être mise en œuvre pour suivre l'évolution de la dépression dans le temps et mieux comprendre les dynamiques individuelles. Enfin, des analyses complémentaires intégrant des variables psychologiques (estime de soi, anxiété, soutien social) offriraient une compréhension plus fine des mécanismes à l'origine de la dépression étudiante.

# TABLEAU DE BORD

## TABLEAU DE BORD DES FACTEURS INFLUENÇANT LA DEPRESSION CHEZ LES ETUDIANTS



## ANNEXE

<<Code source Python>>

# Préparation des données

## Importation

import pandas as pd

import numpy as np

```

import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.impute import KNNImputer
from scipy.stats.mstats import winsorize
file_path = r"C:/Users/DELL/OneDrive/Desktop/INSSEDS/Small Project Stat Inferentielle/Student_Depression.csv"
data = pd.read_csv(file_path)
data
data.dtypes
## Traitement de doublons
data.duplicated().sum()
## Formatage
data = data.drop(['id'], axis = 1)
variable = [
'sexe', 'ville', 'profession', 'duree_sommeil', 'habitudes_alimentaires',
'diplome_suivi', 'pensees_suicidaire',
'antecedants_familiaux_maladie_mentale', 'depression']
for col in variable:
    data[col] = data[col].astype('category')
data['sexe'] = data['sexe'].replace({"Male": "Homme", "Female": "Femme"})
data['pensees_suicidaire'] = data['pensees_suicidaire'].replace({"No": "Non", "Yes": "Oui"})
data['antecedants_familiaux_maladie_mentale'] =
data['antecedants_familiaux_maladie_mentale'].replace({"No": "Non", "Yes": "Oui"})
data['depression'] = data['depression'].replace({0: "Non", 1: "Oui"})
print(data.columns.tolist())
data.columns = data.columns.str.strip()
data.dtypes
## Visualisation des valeurs manquantes
missing_values = data.isnull().sum()

# Visualiser les valeurs manquantes
plt.figure(figsize=(10, 4))

```

```

missing_values.plot(kind='bar', color='black')
plt.title('Nombre de valeurs manquantes par colonne')
plt.xlabel('Colonnes')
plt.ylabel('Nombre de valeurs manquantes')
plt.xticks(rotation=80)
plt.savefig('image1.png')
plt.show()

## Traitement des valeurs manquantes

nbr_missing_values = data.isnull().any(axis = 1).sum()
percentage = nbr_missing_values / len(data)
percentage

data = data.dropna()

missing_values = data.isnull().sum()

# Visualiser les valeurs manquantes

plt.figure(figsize=(10, 4))

missing_values.plot(kind='bar', color='black')
plt.title('Nombre de valeurs manquantes par colonne après traitement')
plt.xlabel('Colonnes')
plt.ylabel('Nombre de valeurs manquantes')
plt.xticks(rotation=80)
plt.savefig('image2.png')
plt.show()

## Visualisation des valeurs aberrantes

num_vars = data.select_dtypes(include=['float64', 'int64']).columns.tolist()

plt.figure(figsize=(12, 8))

for i, var in enumerate(num_vars, 1):

    plt.subplot(3, 3, i)

    sns.boxplot(y=data[var], color='lightblue')

    plt.title(f"{var}")

    plt.tight_layout()

    plt.savefig('image5.png')

```

```

plt.show()

## Traitement des valeurs manquantes

# Winsorisation sur les variables numériques (limiter à 5% de chaque extrême)

for var in num_vars:

    data[var] = winsorize(data[var], limits=[0.05, 0.05])

# Après winsorisation : visualisation par boxplots

plt.figure(figsize=(12, 8))

for i, var in enumerate(num_vars, 1):

    plt.subplot(3, 3, i)

    sns.boxplot(y=data[var], color='lightgreen')

    plt.title(f"{var} après Winsorisation")

plt.tight_layout()

plt.savefig('image6.png')

plt.show()

# Analyse Univariée

## Etude des variables qualitatives

### Paramètres Statistiques

import scipy.stats as stats

# 1. Variables numériques et catégorielles

quantitative_vars_base = data.select_dtypes(include=['float64', 'int64']).columns.tolist()

qualitative_vars_base = data.select_dtypes(include='category').columns.tolist()

# 2. Analyse des variables numériques

print("\n[12/34] Analyse univariée des variables numériques :")

for var in quantitative_vars_base:

    print(f"\n--- {var} ---")

    print(f"Moyenne : {data[var].mean():.2f}")

    print(f"Médiane : {data[var].median():.2f}")

    print(f"Mode : {data[var].mode()[0]:.2f}")

    print(f"Écart-type : {data[var].std():.2f}")

    print(f"Variance : {data[var].var():.2f}")

    print(f"Min : {data[var].min():.2f} | Max : {data[var].max():.2f}")

```

```

print(f"Asymétrie (Skewness) : {stats.skew(data[var]):.2f}")

print(f"Aplatissement (Kurtosis) : {stats.kurtosis(data[var]):.2f}")

# 3. Analyse des variables catégorielles

print("\n📊 Analyse univariée des variables catégorielles :")

for var in qualitative_vars_base:

    print(f"\n--- {var} ---")

    print(data[var].value_counts())

    print("\nProportions :")

    print(data[var].value_counts(normalize=True).round(2))

### Visualisation

# Configuration du style

sns.set_theme(style="whitegrid")

sns.set_palette("husl")

# Séparation des variables qualitatives et quantitatives

qualitative_vars = ['sexe', 'duree_sommeil', 'habitudes_alimentaires', 'diplome_suivi',
'nombre_heure_travail_etude',

    'pensees_suicidaire', 'antecedants_familiaux_maladie_mentale', 'depression']

# Fonction pour afficher le nombre d'individus sur les barres

def annotate_bars(ax):

    for p in ax.patches:

        ax.annotate(f'{int(p.get_height())}',

                    (p.get_x() + p.get_width() / 2., p.get_height()),

                    ha='center', va='center',

                    xytext=(0, 5),

                    textcoords='offset points')

## Visualisation des variables qualitatives

plt.figure(figsize=(20, 15))

plt.suptitle("Analyse univariée des variables qualitatives", y=1.02, fontsize=16)

for i, var in enumerate(qualitative_vars, 1):

    plt.subplot(3, 3, i)

    ax = sns.countplot(x=var, data=data)

```

```

plt.title(f"Distribution de {var}")

plt.xticks(rotation=80)

annotate_bars(ax)

plt.tight_layout()

plt.savefig('image7.png')

plt.show()

## Etude des variables quantitatives

### Visualisation

quantitative_vars = ['age', 'pression_academique', 'moyenne_notes', 'satisfaction_etudes',
'nombre_heure_travail_etude', 'stress_financier']

## Visualisation des variables quantitatives

plt.figure(figsize=(20, 15))

plt.suptitle("Analyse univariée des variables quantitatives", y=1.02, fontsize=16)

for i, var in enumerate(quantitative_vars, 1):

    plt.subplot(3, 3, i)

    # Histogramme avec densité

    sns.histplot(data[var], kde=True, color='green')

    plt.title(f"Distribution de {var}")

    plt.xlabel(var)

    plt.ylabel("Fréquence")

    plt.tight_layout()

    plt.savefig('image8.png')

    plt.show()

# Analyse bivariée

## Normalité

# Sélection des colonnes quantitatives

quant_cols = data.select_dtypes(include=['float64', 'int64']).columns

n_cols = 3 # Nombre de colonnes dans la grille

n_rows = (len(quant_cols) // n_cols + 1) # Calcul automatique des lignes

# Création de la figure

plt.figure(figsize=(18, n_rows * 5)) # Ajustement dynamique de la hauteur

```

```

# Génération des Q-Q Plots dans la grille

for idx, col in enumerate(quant_cols, 1):

    plt.subplot(n_rows, n_cols, idx)

    stats.probplot(data[col].dropna(), dist="norm", plot=plt)

    plt.title(f'Q-Q Plot de {col}', fontsize=12)
    plt.xlabel('Quantiles théoriques', fontsize=10)
    plt.ylabel('Quantiles observés', fontsize=10)

    plt.tight_layout() # Ajustement automatique des espacements

    plt.savefig('image11.png')
    plt.show()

## Pour une variable dquanti et quali

plt.figure(figsize=(8, 6))

sns.boxplot(x='depression', y='age', data=data, palette=['skyblue', 'salmon'])

plt.title('Distribution de l\'Âge par Statut Dépressif')
plt.xlabel('Dépression')
plt.ylabel('Âge')

plt.savefig('image12.png')
plt.show()

## Test de liaison

from scipy.stats import mannwhitneyu

depression_groups = data.groupby('depression') # 'depression' est binaire (0/1)

for col in data.select_dtypes(include=['float64', 'int64']).columns:

    group1 = depression_groups.get_group("Non")[col].dropna() # Groupe sans dépression
    group2 = depression_groups.get_group("Oui")[col].dropna() # Groupe avec dépression

    stat, p = mannwhitneyu(group1, group2)

    print(f'{col} vs dépression: p-value = {p:.4f}')

    print("→ Liaison significative" if p < 0.05 else "→ Pas de liaison significative\n")

## Pour deux variables quali

# Créer un tableau croisé (tableau de contingence) des données

contingency_table = pd.crosstab(data['depression'], data['sexe'])

print(contingency_table)

```

```

# Traçage

contingency_table.plot(kind="bar", stacked=True, figsize=(10,6))

plt.title("Distribution du sexe par Statut Dépressif")

plt.ylabel("Count")

plt.xlabel("depression")

plt.legend(title="sexe")

plt.xticks(rotation=0)

plt.savefig('image21.png')

plt.show()

## Test de liaison

from scipy.stats import chi2_contingency

# Sélection des variables qualitatives (object, category, bool)

qual_cols = data.select_dtypes(include=['object', 'category', 'bool']).columns

depression_col = 'depression' # À adapter si nécessaire

results = []

for col in qual_cols:

    if col != depression_col: # Éviter de tester la variable avec elle-même

        # Création du tableau de contingence

        contingency_table = pd.crosstab(data[depression_col], data[col])

        # Test du chi2

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

        # Vérification de la condition de Cochran

        valid_cells = (expected >= 5).sum()

        total_cells = expected.size

        cochran_ok = (valid_cells / total_cells) >= 0.8 # 80% de cellules valide

        # Stockage des résultats

        results.append({

            'Variable': col,

            'Chi2': chi2,

            'p-value': p,

            '% cellules valides': (valid_cells / total_cells * 100),

```

```

'Cochran': cochrان_ok,
'Degrés liberté': (contingency_table.shape[0]-1)*(contingency_table.shape[1]-1)
}

# Création du dataframe de résultats
results_df = pd.DataFrame(results)
results_df = results_df.sort_values(by='p-value')
# Affichage des résultats
pd.set_option('display.float_format', '{:.3f}'.format)
print("Résultats des tests d'association avec la dépression :")
print(results_df[['Variable', 'Chi2', 'Degrés liberté', 'p-value', '% cellules valides', 'Cochran']])
# Affichage des variables significatives (seuil à 0.05)
print("\nVariables significativement associées (p < 0.05) :")
print(results_df[results_df['p-value'] < 0.05]['Variable'].to_string(index=False))
# Graphique des p-values
plt.figure(figsize=(10, 6))
plt.barh(results_df['Variable'], -np.log10(results_df['p-value']), color='skyblue')
plt.axvline(-np.log10(0.05), color='red', linestyle='--')
plt.title('Significativité des variables qualitatives avec la dépression')
plt.xlabel('p-value')
plt.ylabel('Variables qualitatives')
plt.tight_layout()
plt.savefig('image30.png')
plt.show()

# Statistique Inférentielle
1)
# Supposons que tu as déjà chargé ton dataset
# data = pd.read_csv(...) # si ce n'est pas encore fait

# Étape 1 : encoder la variable en binaire
# Ex : "Oui" → 1, "Non" → 0 (adapte selon tes modalités)
data['pensees_suicidaire_bin'] = data['pensees_suicidaire'].map({'Oui': 1, 'Non': 0}).astype(int)

```

```
# Étape 2 : initialiser le bootstrap
```

```
n_iterations = 1000
```

```
sample_size = len(data)
```

```
print(sample_size)
```

```
bootstrap_props = []
```

```
for i in range(n_iterations):
```

```
    # Échantillon avec remise
```

```
    sample = data['pensees_suicidaire_bin'].sample(n=sample_size, replace=True)
```

```
    # Calcul de la proportion
```

```
    prop = sample.mean()
```

```
    bootstrap_props.append(prop)
```

```
# Étape 3 : calcul de l'intervalle de confiance (IC à 95 %)
```

```
lower_bound = np.percentile(bootstrap_props, 2.5)*100
```

```
upper_bound = np.percentile(bootstrap_props, 97.5)*100
```

```
# Affichage des résultats
```

```
print(f"Intervalle de confiance à 95 % pour la proportion d'étudiants ayant eu des pensées suicidaires : [{lower_bound:.3f}%, {upper_bound:.3f}%]" )
```

2)

## Fonction Bootstrap

```
from tqdm import tqdm # Pour une barre de progression
```

```
def bootstrap_ci(data, stat='mean', n_bootstrap=5000, ci=95):
```

```
    """
```

Calcule l'intervalle de confiance par bootstrap.

Args:

data: Array-like (données à analyser).

stat: 'mean' ou 'median'.

n\_bootstrap: Nombre de rééchantillonnages.

ci: Niveau de confiance (par défaut 95%).

Returns:

(statistique, IC\_bas, IC\_haut)

```
    """
```

```

stats = []

for _ in tqdm(range(n_bootstrap)):

    sample = np.random.choice(data, size=len(data), replace=True)

    if stat == 'mean':

        stats.append(np.mean(sample))

    elif stat == 'median':

        stats.append(np.median(sample))

lower = np.percentile(stats, (100 - ci) / 2)

upper = np.percentile(stats, ci + (100 - ci) / 2

if stat == 'mean':

    return np.mean(data), lower, upper

elif stat == 'median':

    return np.median(data), lower, upper

## Intervalle de confiance

# Filtrer les étudiants dépressifs

depressifs = data[data['depression'] == 'Oui']

heures_travail = depressifs['nombre_heure_travail_etude'].dropna() # Supprimer les NaN

# Bootstrap pour la moyenne

moyenne, ci_low_moy, ci_high_moy = bootstrap_ci(heures_travail, stat='mean')

print(f"**Heures de travail/études (Dépressifs)**")

print(f"**Moyenne (IC 95%): {moyenne:.1f} heures [{ci_low_moy:.1f}, {ci_high_moy:.1f}]**")

# Bootstrap pour la médiane

mediane, ci_low_med, ci_high_med = bootstrap_ci(heures_travail, stat='median')

print(f"**Médiane (IC 95%): {mediane:.1f} heures [{ci_low_med:.1f}, {ci_high_med:.1f}]**\n")

3)

# Séparation des groupes

stress_depressifs = data[data['depression'] == 'Oui']['stress_financier'].dropna()

stress_non_depressifs = data[data['depression'] == 'Non']['stress_financier'].dropna()

# Fonction pour afficher les résultats

def print_results(stress_data, groupe):

    moyenne, ci_low_moy, ci_high_moy = bootstrap_ci(stress_data, stat='mean')

```

```

mediane, ci_low_med, ci_high_med = bootstrap_ci(stress_data, stat='median')

print(f"**Stress financier ({groupe}**")

print(f"Moyenne (IC 95%): {moyenne:.1f} [{ci_low_moy:.1f}, {ci_high_moy:.1f}]")

print(f"Médiane (IC 95%): {mediane:.1f} [{ci_low_med:.1f}, {ci_high_med:.1f}]\n")

# Résultats

print_results(stress_depressifs, "Avec Dépression")
print_results(stress_non_depressifs, "Sans Dépression")

4)

# Importations

from scipy.stats import mannwhitneyu

from tqdm import tqdm

# 1. Préparation des données

depressifs = data[data['depression'] == 'Oui']['satisfaction_etudes'].dropna()

non_depressifs = data[data['depression'] == 'Non']['satisfaction_etudes'].dropna()

# 2. Test de Mann-Whitney U

stat, p_value = mannwhitneyu(depressifs, non_depressifs, alternative='two-sided')

print(f"Test de Mann-Whitney U : p-value = {p_value:.4f}")

# Interprétation

alpha = 0.05

if p_value < alpha:

    print("→ Différence significative (rejet de H0)")

else:

    print("→ Pas de différence significative (échec pour rejeter H0)")

5)

Aucune différence car les données de la variable satisfaction_ travail sont constantes et égale à zéro

6)

# Nettoyage éventuel des colonnes (espace, majuscules...)

data.columns = data.columns.str.strip()

# Création de la table de contingence

table = pd.crosstab(data['depression'], data['habitudes_alimentaires'])

```

```

# Affichage de la table
print("Table de contingence :\n", table)

# Test du chi2
chi2, p, dof, expected = stats.chi2_contingency(table)

# Résultats
print(f"\nStatistique de chi2 : {chi2:.4f}")
print(f"p-value : {p:.4f}")

# Interprétation
alpha = 0.05
if p < alpha:
    print("☒ Rejet de H0 : la dépression dépend des habitudes alimentaires.")
else:
    print("☑ On ne rejette pas H0 : la dépression est indépendante des habitudes alimentaires.")

# Création de la heatmap
plt.figure(figsize=(8, 5))

sns.heatmap(table, annot=True, fmt="d", cmap="YlGnBu", cbar=True)
plt.title("Relation entre la dépression et les habitudes alimentaires")
plt.xlabel("Habitudes alimentaires")
plt.ylabel("Dépression")
plt.tight_layout()
plt.savefig('image31.png')
plt.show()

7)

# Nettoyage éventuel des noms de colonnes
data.columns = data.columns.str.strip()

# Table de contingence
table = pd.crosstab(data['depression'], data['duree_sommeil'])

# Affichage de la table
print("Table de contingence :\n", table)

# Test du Chi2
chi2, p, dof, expected = stats.chi2_contingency(table)

```

```
print(f"\nStatistique de chi2 : {chi2:.4f}")
print(f"p-value : {p:.4f}")

# Interprétation
alpha = 0.05
if p < alpha:
    print("☒ Rejet de H0 : la dépression dépend de la durée du sommeil.")
else:
    print("☑ On ne rejette pas H0 : la dépression est indépendante de la durée du sommeil.")

# Optionnel : Heatmap
plt.figure(figsize=(8, 5))
sns.heatmap(table, annot=True, fmt="d", cmap="YlOrBr", cbar=True)
plt.title("Relation entre dépression et durée du sommeil")
plt.xlabel("Durée du sommeil")
plt.ylabel("Dépression")
plt.xticks(rotation=10)
plt.tight_layout()
plt.savefig('image32.png')
plt.show()
```