MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

RÉPUBLIQUE DE CÔTE D'IVOIRE

Institut Supérieur de Statistique
d'Économétrie et de Data Science

MASTER II

STATISTIQUE – ÉCONOMÉTRIE – DATA SCIENCE

Mini-projet

Statistique Inférentielle

ÉTUDE DES FACTEURS INFLUENÇANT LA DÉPRESSION CHEZ LES ÉTUDIANTS

2024 – 2025

Étudiant

ABBE JEAN PIERRE

Enseignant – Encadreur

AKPOSSO DIDIER MARTIAL

AVANT-PROPOS

Dans le cadre de la formation du **MASTER II** en **Statistique-Économétrie-Data Sciences**, à l'INSSEDS (Institut Supérieur de Statistique d'Économétrie et de Data Science), l'une des **exigences** conditionnant la **validation** des **différents crédits** est la réalisation d'un **mini-projet** à la **fin** de chaque **module** de la formation. C'est dans ce cadre académique que s'inscrit la présente étude.

Par ailleurs, **ce mini-projet est relatif à l'analyse statistique inférentielle** dans lequel nous répondrons uniquement à un certain nombre de questions déjà répertoriées. Ceci étant, en tant qu'auteur, nous n'avons **pas la prétention** d'avoir **abordé** tous les **contours nécessaires** pour **y revêtir** le caractère d'une **véritable étude thématique**.

Également, toutes les **conclusions** développées dans ce **présent rapport relèvent uniquement de l'auteur** et **n'engagent ni autrui, ni l'INSSEDS** (Institut Supérieur de Statistique d'Économétrie et de Data Science).

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES GRAPHIQUES

INTRODUCTION GÉNÉRALE

Dans un contexte académique où le bien-être des étudiants revêt une importance cruciale, la compréhension des facteurs influençant la **dépression chez les étudiants** devient essentielle. La dépression chez les jeunes adultes peut être exacerbée par de nombreux éléments, tels que la pression académique, le stress lié aux études, les habitudes de sommeil, et les facteurs sociaux ou familiaux. Identifier ces facteurs et comprendre leur impact sur la santé mentale des étudiants est crucial pour mettre en place des interventions adaptées visant à prévenir ou réduire les risques de dépression.

La problématique centrale est alors de savoir : **comment analyser les différents facteurs qui influencent la dépression chez les étudiants pour proposer des stratégies de soutien et d'intervention efficaces ?**

Pour répondre à cette problématique, nous utiliserons des techniques statistiques avancées, telles que les **tests d'hypothèses**, **l'analyse de variance (ANOVA)**, et les **tests de corrélation**, afin d'explorer les relations entre la dépression et des variables telles que la pression académique, la durée du sommeil, le stress financier, et d'autres facteurs démographiques. L'objectif est de mieux comprendre ces facteurs afin de pouvoir proposer des solutions pratiques aux établissements éducatifs pour améliorer la santé mentale des étudiants.

Les résultats attendus de cette étude incluent :

- Une **analyse descriptive** des principales variables influençant la dépression (pression académique, durée du sommeil, etc.).
- L'identification des **facteurs significatifs** contribuant à la dépression chez les étudiants.
- Des recommandations sur les **mesures préventives** et les stratégies de soutien adaptées.

Cette étude est structurée en quatre parties : dans la **première partie**, **nous analysons et préparons les données**. Les **trois autres parties suivantes** sont dédiées à **l'analyse des relations entre la dépression et les variables étudiées à l'aide de l'estimation et des tests statistiques**. Enfin, la **dernière partie** présente des **conclusions et** des **recommandations** pour les interventions basées sur les résultats obtenus.

1ère PARTIE : APPROCHE MÉTHODOLOGIQUE DES DONNÉES

Cette partie présente les données utilisées et l'analyse descriptive des données.
La méthodologie de notre étude est basée sur une approche utilisant le logiciel Python.

1. Présentation du jeu de données initiale

Notre jeu de données initiale nommé «**Student_Depression**» **comporte 27.901 observations et 18 variables.** Ces variables sont listées dans le tableau de dictionnaire de données ci-dessous :

VARIABLE	TYPE DE DE DONNÉES
ID	int64
SEXE	object
AGE	float64
VILLE	object
PROFESSION	object
PRESSION_ACADEMIQUE	float64
PRESSION_LIEE_AU_TRAVAIL	float64
MOYENNE_NOTES	float64
SATISFACTION_ETUDES	float64
SATISFACTION_TRAVAIL	float64
DUREE_SOMMEIL	object
HABITUDES_ALIMENTAIRES	object
DIPLOME_SUIVI	object
PENSEES_SUICIDAIRE	object
NOMBRE_HEURE_TRAVAIL_ETUDE	float64
STRESS_FINANCIER	float64
ANTECEDANTS_FAMILIAUX_MALADIE_MENTALE	object
DEPRESSION	int64

Tableau 1 : Tableau du dictionnaire de données

2. Recodage des variables

Le recodage des variables est une étape essentielle dans le traitement des données. Il permet de transformer les variables brutes en formats adaptés pour l'analyse statistique. Dans notre cas, les identifiants ont été convertis en chaînes de caractères, et les variables binaires telles que **'sexe'**, **'pensees_suicidaire'**, et **'depression'** ont été recodées en catégories explicites comme **'Femme'** et **'Homme'**, ou **'Oui'** et **'Non'**. Ce recodage a facilité la compréhension et l'interprétation des résultats. Les ***variables catégorielles ont*** également ***été converties en types «category»***, pour l'optimisation de la gestion de la mémoire afin de simplifier les tests statistiques, notamment le **test du Khi-deux**. **En structurant correctement les données, le recodage a permis d'éviter les erreurs liées à des types incompatibles, garantissant ainsi des analyses fiables et cohérentes.**

3. Apurement des valeurs manquantes

Cette analyse a d'abord démarré par un recodage des variables qualitatives binaires, après quoi avons-nous identifié **3 individus ayant une information manquante**, soit **moins de 1% du total**. Puisque cette **proportion** est **négligeable**, **nous avons procédé à la suppression de ces individus ayant des valeurs manquantes du jeu de données**.

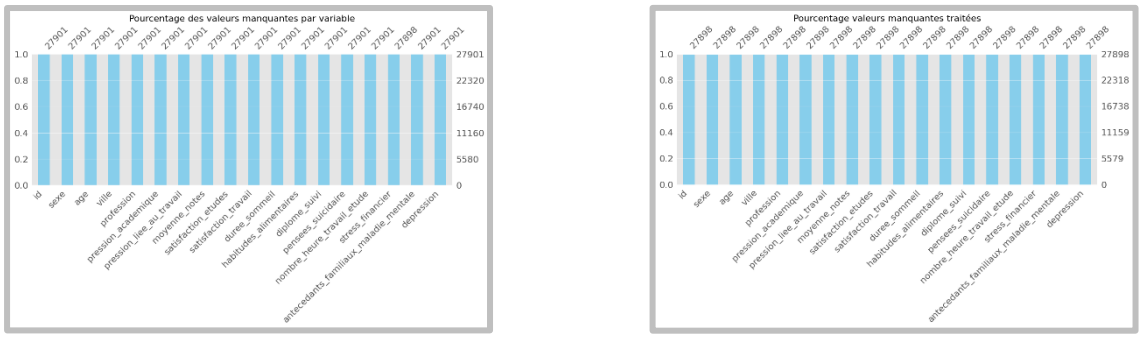


Figure 1 : Visualisation des données avant et après le traitement

4. Traitement des valeurs aberrantes et/ou extrêmes

En plus du traitement des valeurs manquantes, nous traiterons les valeurs aberrantes et/ou extrêmes. **L'exploration du jeu de données nous a permis d'identifier des valeurs aberrantes et/ou extrêmes**. Cela est perceptible au travers des **débordements observés sur les boites à moustaches ci-dessous**. La technique de Winzorisation a été utilisée pour traiter ces valeurs aberrantes et/ou extrêmes en les ramenant dans les limites des bornes (inférieure et supérieure).

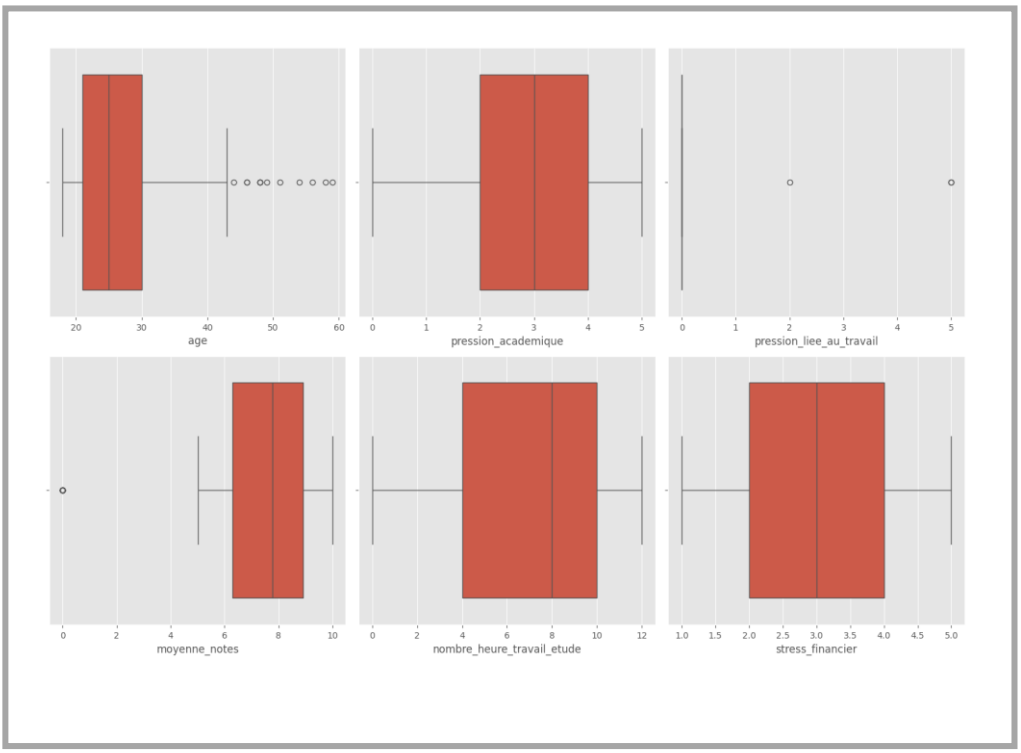


Figure 3 : Boxplots avant traitement

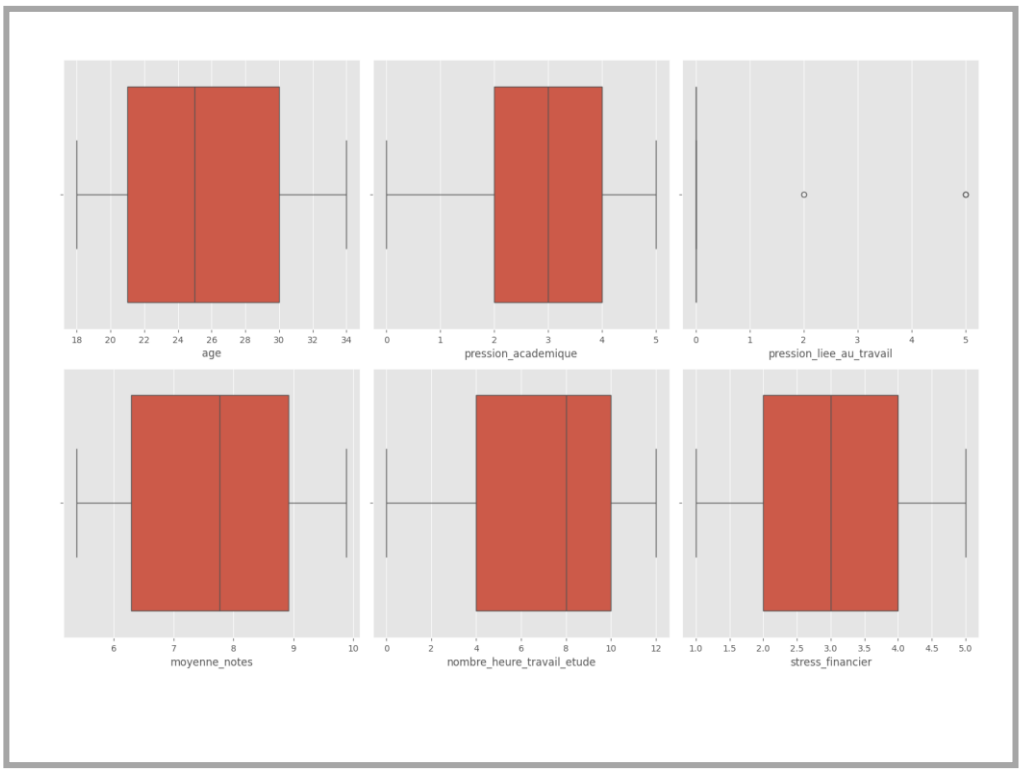


Figure 2 : Boxplots après traitement

2ème PARTIE : ESTIMATION STATISTIQUE

Cette partie se consacre exclusivement à l'estimation statistique des variables prises individuellement, en calculant des mesures descriptives telles que la moyenne, la médiane, et en utilisant des tests statistiques et des intervalles de confiance.

I. Estimation de la proportion des étudiants ayant déjà eu des pensées suicidaires

1. Tableau statistique

L'objectif de cette analyse est de déterminer l'intervalle de confiance à 95 % pour la proportion d'étudiants ayant rapporté avoir déjà eu des pensées suicidaires. Cette estimation se base sur un échantillon de 27,898 étudiants, parmi lesquels 17,656 ont répondu affirmativement (c'est-à-dire qu'ils ont indiqué avoir eu des pensées suicidaires), tandis que 10,242 ont répondu négativement.

	Effectif	Fréquence
Non	10,242	0,36712
Oui	17,656	0,63288

Tableau 2 : Tableau de la proportion des étudiants selon la pensée suicidaire

Sans estimation, la proportion d'étudiants ayant eu des pensées suicidaires 17,656/27,898=0,63288 ou **63,29%.** Or, cette estimation ponctuelle de 63,29 % peut être biaisée par des erreurs d'échantillonnage, c'est pourquoi un intervalle est calculé pour exprimer la **plage probable** dans laquelle la proportion réelle pourrait se trouver dans la population. Ainsi donc, nous ferons un test binomial afin de vérifier la validité de l'estimation de la proportion d'étudiants ayant des pensées suicidaires.

2. Estimation de la proportion avec intervalle de confiance (test binomial)

Le test binomial a été utilisé pour estimer la proportion d'étudiants ayant répondu "Oui" à la question concernant les pensées suicidaires, et pour obtenir l'intervalle de confiance associé à cette proportion. Le test binomial est particulièrement adapté ici car il permet d'estimer la proportion d'événements dans un échantillon binaire (ici "Oui" ou "Non") et de fournir un intervalle de confiance pour cette proportion.

L'intervalle de confiance à 95 % permet de déterminer l'intervalle dans lequel nous pouvons être à 95 % confiants que la proportion réelle d'étudiants ayant eu des pensées suicidaires se situe dans la population étudiée.

Avec un niveau de confiance de 95 % que la proportion réelle d'étudiants ayant eu des pensées suicidaires dans la population étudiée **se situe entre 62,7 %** et **63,9 %**

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

1. Tableau statistique du sous-ensemble d'étudiants dépressifs

Un filtre a été appliqué au jeu de données **pour isoler les étudiants souffrant de dépression** (depression == "Oui"). Cela a permis de créer un sous-ensemble spécifique contenant uniquement les données pertinentes pour cette analyse, notamment la variable "nombre d'heures de travail ou d'études". **Cette étape garantit que les calculs ultérieurs se concentrent uniquement sur la population cible**.

N.B : Vu la taille du tableau (16.335 lignes), nous ne l'afficherons pas ici !

2. Test de normalité des données

Le test de Shapiro-Wilk a été utilisé pour vérifier si la distribution des heures de travail ou d'études chez les étudiants dépressifs suit une loi normale. Les résultats ont donné **une statistique W = 0,9** avec une **p-value < $2,2^{e-16}$**, indiquant que **la distribution est significativement différente d'une distribution normale**.

3. Estimation de la moyenne par bootstrap

Le bootstrap (1 000 réplications) **a permis d'estimer la moyenne des heures de travail ou d'études chez les étudiants dépressifs**. **La moyenne calculée** est de **7,81 heures**, avec un **intervalle de confiance à 95 %** entre **7,76 et 7,86 heures**. Cette méthode est appropriée pour des données non normales, car elle repose sur des échantillons répétés pour obtenir des estimations robustes.

4. Estimation de la médiane par bootstrap

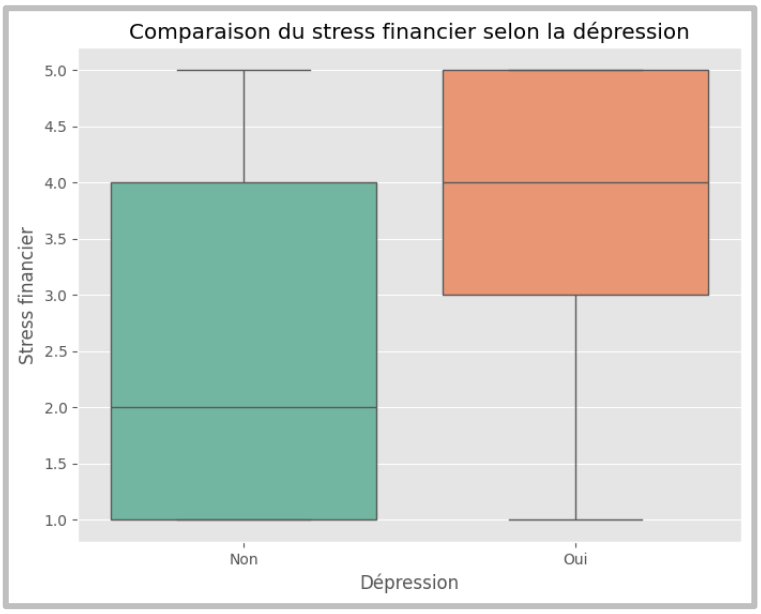
Le même procédé de bootstrap (1 000 réplications) a été utilisé pour estimer la médiane, qui est de **9 heures**, avec un intervalle de confiance à 95 % également fixé à **9 heures**. Ces résultats suggèrent que **la moitié des étudiants dépressifs passent au moins 9 heures par jour à travailler ou étudier**, ce qui reflète la centralité de cette valeur dans la distribution.

Tous ces résultats indiquent que **les étudiants souffrant de dépression passent en moyenne 7,81 heures** par jour à travailler ou étudier, mais **la médiane** de **9 heures montre que la moitié d'entre eux travaille ou étudie au moins 9 heures par jour**. La différence entre la moyenne et la médiane suggère que la distribution des heures de travail ou d'études est asymétrique, probablement influencée par des valeurs plus faibles.

III. Évaluons la moyenne et la médiane du stress financier pour les étudiants avec et sans dépression

Évaluer la moyenne et la médiane du stress financier pour les étudiants avec et sans dépression implique une comparaison entre deux populations : les étudiants souffrant de dépression et ceux qui n'en souffrent pas. ***Cela peut être vu*** *comme une* ***comparaison de populations***, notamment dans le cadre de la distribution du stress financier. Nous suivrons donc les étapes suivantes *sans faire de tests d'hypothèses*.

1. Comparaison graphique des deux populations



Graphique 1 : Comparaison du stress financier selon la dépression

Avant tout, nous avons séparé les étudiants en deux groupes : ceux souffrant de dépression et ceux n'en souffrant pas. Le **graphique 1** visualise la différence de stress financier entre les étudiants souffrant de dépression et ceux qui ne le sont pas. Il montre une plus grande dispersion et des valeurs plus élevées de stress financier chez les étudiants dépressifs.
On peut soupçonner une différence significative entre les deux populations. Toutefois, à l'aide de test d'hypothèses nous pourrons confirmer ou infirmer ces soupçons (sauf que nous ne les ferrons pas) !

2. Test de normalité des données (sur les données du stress financier pour les étudiants non dépressifs)

Le test de Shapiro-Wilk a été utilisé pour vérifier si la distribution suit une loi normale. Et il a montré que la distribution n'est pas normale car la **p-value < $2{,}2^{e\text{-}16}$**.

3. **<u>Estimation des moyennes par bootstrap</u>**

- Pour les **étudiants dépressifs**, la **moyenne du stress financier** est estimée à **3,58** avec un **intervalle de confiance à 95% de [3,56 ; 3,60]**.
- Pour les **étudiants non dépressifs**, la **moyenne est estimée** à **2,52** avec un **intervalle de confiance à 95% de [2,49 ; 2,54]**.

4. **<u>Estimation des médianes par bootstrap</u>**

- Pour les **étudiants dépressifs**, la **médiane du stress financier** est **4**, avec un **intervalle de confiance de [4 ; 4]**.
- Pour les **étudiants non dépressifs**, la **médiane est 2,** avec un **intervalle de confiance de [2 ; 2]**.

Les résultats montrent que **les étudiants souffrant de dépression ont une moyenne et une médiane plus élevées en termes de stress financier par rapport à ceux n'en souffrant pas**. La différence de moyenne (3,58 vs 2,52) et de médiane (4 vs 2) suggère que *la dépression pourrait être associée à un stress financier plus important.*

3ème PARTIE : TEST DE COMPARAISON DE POPULTIONS

Cette partie est consacrée aux tests de différences de moyennes et de médianes entre populations, en en **effectuant des tests statistiques pour comparer les moyennes/ médianes de deux groupes indépendants** sur une variable d'intérêt. Cela permet de déterminer si les différences observées entre les moyennes/ médianes des deux groupes sont statistiquement significatives ou si elles pourraient être dues au hasard.

I. LA SATISFACTION DES ÉTUDES DIFFÈRE-T-ELLE SIGNIFICATIVEMENT ENTRE LES ÉTUDIANTS SOUFFRANT DE DÉPRESSION ET CEUX QUI N'EN SOUFFRENT PAS ?

1. Test de comparaison graphique

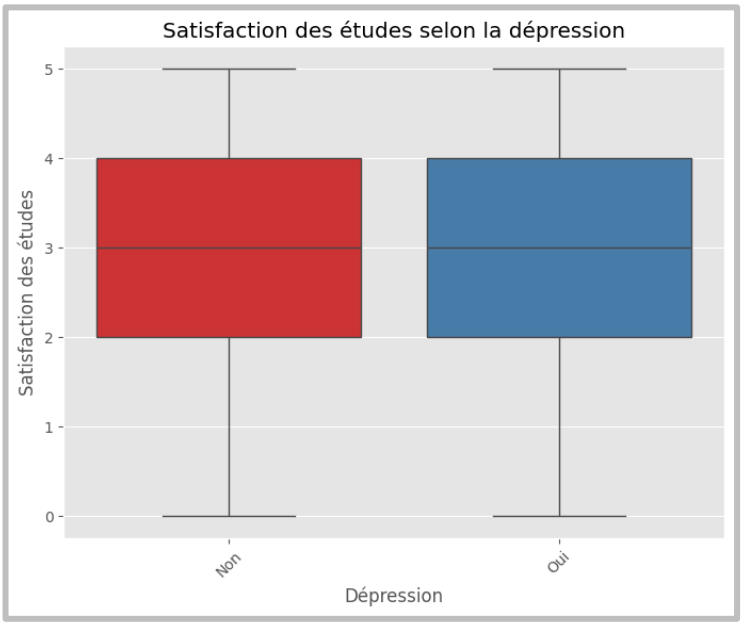
Dans l'ensemble, 0,0143% des étudiants non dépressifs ont répondu qu'ils sont très insatisfaits (score 0), tandis que 0,0215% des étudiants dépressifs ont donné la même réponse. En ce qui concerne la satisfaction générale, 5,71% des étudiants non dépressifs et 13,82% des étudiants dépressifs ont évalué leur satisfaction à 1. De même, 7,42% des étudiants non dépressifs et 13,51% des étudiants dépressifs ont répondu avec un score de 2.

En revanche, pour les scores 3 et 4, la proportion d'étudiants non dépressifs est légèrement plus élevée que celle des étudiants dépressifs, avec des pourcentages respectifs de 8,84% contre 12,02% pour le score 3, et 11,09% contre 11,70% pour le score 4. Enfin, 8,37% des étudiants non dépressifs et 7,48% des étudiants dépressifs ont un score de 5.

Ces résultats suggèrent qu'il y a des variations dans les réponses des deux groupes, mais il n'est pas encore possible de conclure que les différences sont significatives sans effectuer une analyse plus approfondie surtout que le **graphique 2** (voir ci-dessous) ne montre pas de différence significative entre les deux populations.

Satisfaction des études	***Oui***		***Non***	
	Effectif	*Fréquence*	*Effectif*	*Fréquence*
0	4	0,000143	6	0,000215
1	1593	0,057101	3856	0,138218
2	2070	0,074199	3768	0,135063
3	2467	0,088429	3353	0,120188
4	3095	0,11094	3264	0,116998
5	2334	0,083662	2088	0,074844

Tableau 3 : Tableau des effectifs de la satisfaction des études selon la dépression



Graphique 2 : Satisfaction des études selon la dépression

L'observation du graphique 2 laisse très peu entrevoir les différences entre les deux populations. Toutefois par calcul, on trouve :

- **On trouve respectivement une moyenne de 3,22 chez les étudiants non souffrants et 2,75 chez ceux souffrant de dépression.**
- En comparant les quartiles des deux populations, on peut constater que leurs médianes sont égales 3.
- Les **écarts types** pour les scores de relation chez **étudiants non souffrant chez ceux souffrant de dépression** sont respectivement de **1,33 et 1,35**.

Nous allons maintenant vérifier la significativité de ce constat à l'aide de tests statistiques.

2. Test d'hypothèse

Il s'agit ici de faire un **test de comparaison de moyenne de deux populations indépendantes**. Le test approprié est le **test de Student d'égalité de deux moyennes si** sa condition d'utilisation est respectée à savoir **la normalité des données** dans chaque sous population, ajouté à cela l'**égalité des variances**. **Dans le cas contraire nous ferons un test non paramétrique de Wilcoxon ou de Kruskal-Wallis**.

➢ Test de la normalité des données dans chaque sous population

H_0 : La distribution des scores de satisfaction des études suit la loi normale chez les étudiants souffrant de dépression et chez ceux qui n'en souffrent pas.

H_1 : La distribution des scores de satisfaction des études ne suit pas la loi normale chez les étudiants souffrant de dépression et chez ceux qui n'en souffrent pas.

Le test de Shapiro-Wilk pour tester la normalité des données de la satisfaction des études dans chaque groupe (étudiants souffrant de dépression et ceux n'en souffrant pas) donne **des p-values** du test de Shapiro-Wilk pour les deux groupes (étudiants souffrant de dépression et ceux n'en souffrant pas) sont **très inférieures à 0,05 (p-value < $2{,}2^{e-16}$**pour les deux groupes), **nous rejetons l'hypothèse nulle (H_o)** selon laquelle la distribution des scores de satisfaction des études suit la loi normale dans les deux populations. **Cela signifie que la distribution des scores de satisfaction des études ne suit pas la loi normale, tant chez les étudiants souffrant de dépression que chez ceux n'en souffrant pas**. Il serait plus approprié d'utiliser un test non paramétrique, comme le test de Wilcoxon-Mann-Whitney.

- **<u>Test d'égalité des variances</u>**

H_o : $\sigma_1 = \sigma_2$: les variances ne sont pas significativement chez les étudiants non souffrants et ceux souffrant de dépression

H_1 : $\sigma_1 \neq \sigma_2$: les variances sont significativement différentes chez les étudiants non souffrants et ceux souffrant de dépression

La p-value du test F pour comparer les variances est de **0,03303**, ce qui est **inférieur à 0,05**. Par conséquent, **nous rejetons l'hypothèse nulle (H_o) selon laquelle les variances de la satisfaction des études sont égales entre les étudiants souffrant de dépression et ceux n'en souffrant pas**. Cela indique que les variances des deux groupes diffèrent significativement.

- **<u>Test d'égalité des moyennes</u>** : test de Wilcoxon-Mann-Whitney

H_o : $\mu_1 = \mu_2$: la moyenne des scores de satisfaction des études ne diffère pas entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas.

H_1 : $\mu_1 \neq \mu_2$: la moyenne scores de satisfaction des études diffère entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas.

La p-value du test de **Wilcoxon-Mann-Whitney est extrêmement faible (p = $1{,}18^{e-173}$), bien inférieure à 0,05**. Par conséquent, **nous rejetons l'hypothèse nulle (H_o) selon laquelle la satisfaction des études ne diffère pas entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas**.

Cela indique que **la satisfaction des études diffère significativement entre les deux groupes**.

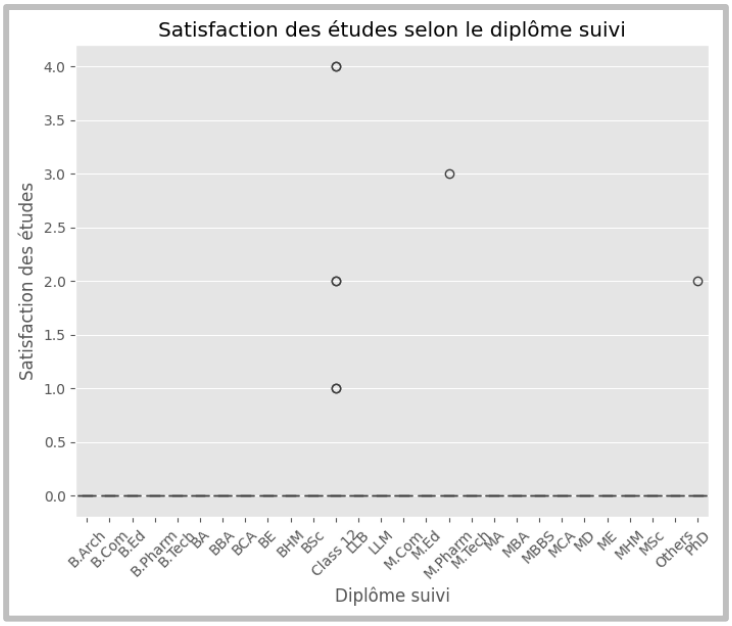
II. LES NIVEAUX DE SATISFACTION AU TRAVAIL DIFFÈRENT-ILS SIGNIFICATIVEMENT SELON LE DIPLOME SUIVI ?

1. Test de comparaison graphique

Dans l'ensemble, nous avons une **domination de l'insatisfaction** (score 0) dans la plupart des catégories de diplômes, à l'exception de **Class 12** et **PhD**, où une certaine **variabilité de satisfaction** est observée. Cette variabilité pourrait suggérer une **diversité d'opinions** sur la satisfaction au travail dans ces groupes. Les autres groupes semblent avoir une perception **plus uniforme**, avec une grande partie des réponses étant concentrées sur l'absence de satisfaction (score 0).

diplome_suivi	*satisfaction_travail*				
	0	**1**	**2**	**3**	**4**
B.Arch	1478	0	0	0	0
B.Com	1506	0	0	0	0
B.Ed	1866	0	0	0	0
B.Pharm	810	0	0	0	0
B.Tech	1152	0	0	0	0
BA	600	0	0	0	0
BBA	696	0	0	0	0
BCA	1432	0	0	0	0
BE	613	0	0	0	0
BHM	925	0	0	0	0
BSc	888	0	0	0	0
MHM	191	0	0	0	0
MSc	1190	0	0	0	0
Others	35	0	0	0	0
Class 12	6074	2	2	0	2
LLB	671	0	0	0	0
LLM	482	0	0	0	0
M.Com	734	0	0	0	0
M.Ed	821	0	0	0	0
M.Pharm	581	0	0	1	0
M.Tech	1022	0	0	0	0
MA	544	0	0	0	0
MBA	562	0	0	0	0
MBBS	695	0	0	0	0
MCA	1044	0	0	0	0
MD	572	0	0	0	0
ME	185	0	0	0	0
PhD	521	0	1	0	0

Tableau 4 : Tableau des effectifs de la satisfaction au travail en fonction du diplôme suivi



Graphique 3 : Satisfaction des études selon le diplôme suivi

Le **graphique 3** représente la **satisfaction au travail en fonction du diplôme suivi**, sous forme de points pour chaque combinaison diplôme-satisfaction :

- La majorité des diplômes (comme **B.Arch**, **B.Com**, **B.Ed**, **B.Tech**, etc.) sont associés à une **satisfaction au travail de niveau 0**. Cela indique une insatisfaction généralisée ou un manque de variabilité dans les réponses pour ces groupes. Cette ***prépondérance des scores de satisfaction à 0*** dans la majorité des diplômes ***pourrait refléter un manque de corrélation entre le diplôme obtenu et la satisfaction au travail*** ;
- Cette uniformité pourrait traduire une perception homogène (négative) parmi les diplômés de ces filières ***(des défis communs affectent les diplômés de toutes les filières dans la satisfaction au travail)*** ;
- Les points isolés pour **Class 12**, **M.Pharm**, et **PhD** montrent que ces groupes incluent des individus ayant une opinion divergente sur leur satisfaction au travail.

Nous allons maintenant vérifier la significativité de **ce constat** à l'aide de tests statistiques.

2. Test d'hypothèse

Il s'agit ici de faire un **test de comparaison de moyenne de deux populations indépendantes**. Le test approprié est le **test de Student d'égalité de deux moyennes si** sa condition d'utilisation est respectée à savoir **la normalité des données** dans chaque sous population, ajouté à cela l'**égalité des variances**. **Dans le cas contraire nous ferons un test non paramétrique de Wilcoxon ou de Kruskal-Wallis**.

- **Test de la normalité des données dans chaque sous population**

H_o : La distribution de la satisfaction au travail est normale pour chaque sous-groupe de diplôme suivi.

H_1 : La distribution de la satisfaction au travail ne suit pas la loi normale pour chaque sous-groupe de diplôme suivi.

Pour la majorité des groupes de diplômes, le test retourne **NA**, ce qui signifie que :

- Les groupes ont trop peu d'observations (moins de 3).
- Les données dans ces groupes sont identiques (pas de variabilité).

Pour certains diplômes, comme **Class 12**, **M.Pharm**, et **PhD**, des p-valeurs significatives ont été obtenues :

- **Class 12**: $\mathrm{p} \approx 1{,}22 \times 10^{-95}$
- **M.Pharm**: $\mathrm{p} \approx 4{,}15 \times 10^{-47}$
- **PhD**: $\mathrm{p} \approx 3{,}48 \times 10^{-45}$

Ces p-valeurs très faibles indiquent que pour ces groupes, **la distribution de la satisfaction au travail n'est pas normale**.

- **Test d'égalité des variances**

H_o : $\sigma_1 = \sigma_2$: les variances de satisfaction au travail sont homogènes entre les groupes

H_1 : $\sigma_1 \neq \sigma_2$: les variances de satisfaction au travail ne sont pas homogènes entre les groupes

La p-value du test F pour comparer les variances est de **0,7613**, ce qui est **supérieur à 0,05**. Par conséquent, **nous acceptons l'hypothèse nulle (H_o) : les variances de satisfaction au travail sont homogènes entre les groupes**.

- **Test de Kruskal-Wallis**

H_o : $\mu_1 = \mu_2$: Toutes les médianes sont égales.

H_1 : $\mu_1 \neq \mu_2$: Au moins un groupe a une médiane différente.

Résultat :

- **χ^2 = 25,327, df = 27, p = 0,5561.**
- **Une p-valeur élevée (p > 0, 05) indique qu'il n'y a pas de différence significative entre les médianes de satisfaction au travail des différents groupes de diplôme.**

N.B : Le test de Wilcoxon compare uniquement **deux groupes indépendants**, tandis que notre analyse inclut **28 groupes de diplômes**. Le test de Kruskal-Wallis est plus approprié ici car il permet une comparaison globale entre plusieurs groupes. Utiliser Wilcoxon impliquerait des **comparaisons multiples** (378 paires possibles), augmentant le risque d'erreurs et rendant l'analyse complexe. De plus, le Kruskal-Wallis est adapté aux données non normales et répond à notre objectif d'évaluer les différences globales entre groupes.

4ème PARTIE : TEST D'INDÉPENDANCE

La partie du **test d'indépendance** est consacrée à évaluer si deux variables qualitatives sont statistiquement indépendantes ou non. Il s'agira de vérifier si **la dépression est liée aux habitudes alimentaires** et si **la durée du sommeil influence la dépression**. Ces analyses permettent d'identifier des relations significatives entre les variables, autrement dit, de savoir si un lien existe ou si ces variables sont indépendantes.

I. LA DÉPRESSION EST-ELLE INDÉPENDANTE DES HABITUDES ALIMENTAIRES ?

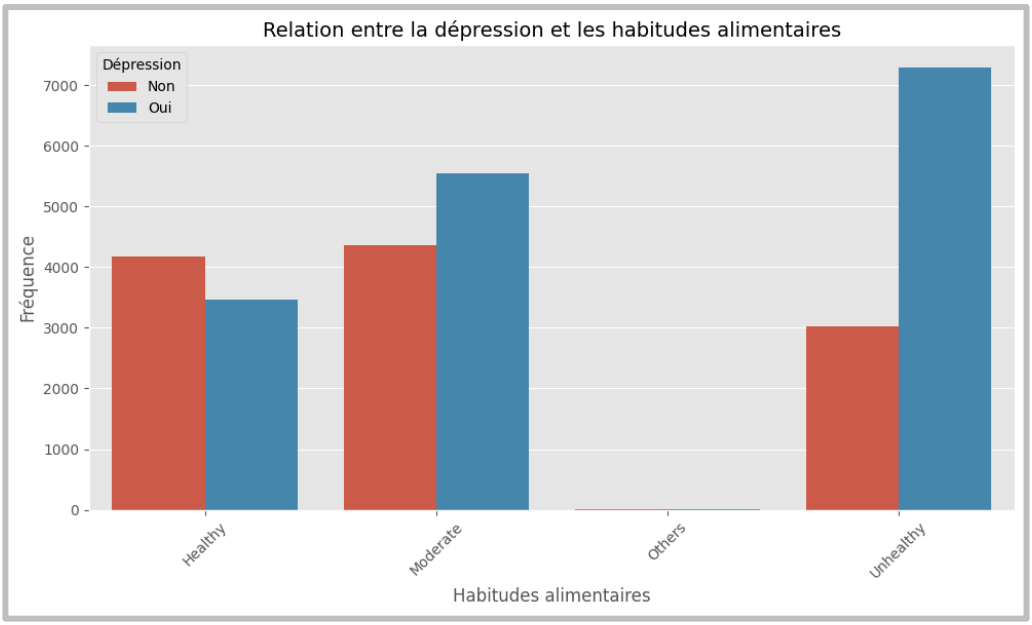
1. Test de comparaison graphique

Les données montrent que la dépression est plus fréquente chez les personnes ayant des habitudes alimentaires malsaines (71%) et modérées (56%). En revanche, elle est moins présente chez celles ayant des habitudes alimentaires saines (45%). Les habitudes alimentaires "autres" sont très peu représentées, mais la dépression y est présente dans 66% des cas. En conclusion, les habitudes alimentaires malsaines semblent être associées à un risque plus élevé de dépression. Cependant, des analyses statistiques supplémentaires seraient nécessaires pour confirmer cette tendance.

Ces résultats suggèrent qu'il y a des variations dans les réponses des deux groupes, mais il n'est pas encore possible de conclure que les différences sont significatives sans effectuer une analyse plus approfondie surtout que le **graphique 2** (voir ci-dessous) ne montre pas de différence significative entre les deux populations.

Dépression	*État de santé*	*Effectif*
Non	Healthy	4177
Oui	Healthy	3472
Non	Moderate	4363
Oui	Moderate	5558
Non	Others	4
Oui	Others	8
Non	Unhealthy	3019
Oui	Unhealthy	7297

Tableau 5 : Tableau des effectifs des habitudes alimentaires selon la dépression



Graphique 4 : Relation entre la dépression et les habitudes alimentaires

Le diagramme en barres montre que dans la catégorie **Healthy**, plus de personnes n'ont pas de dépression (4177) que celles qui en ont (3472). Dans la catégorie **Moderate**, la dépression est plus fréquente (5558 avec dépression contre 4363 sans). Les effectifs dans **Others** sont très faibles (4 sans dépression, 8 avec dépression), ce qui rend cette catégorie peu significative. En revanche, dans la catégorie **Unhealthy**, il y a une forte proportion de personnes avec dépression (7297 contre 3019 sans), suggérant une relation forte entre de mauvaises habitudes alimentaires et la dépression.

2. Test d'hypothèse

Il s'agit ici d'effectuer un **test du Khi-deux** pour évaluer l'indépendance entre la durée du sommeil et la dépression. L'objectif de ce test est de vérifier si les proportions de participants souffrant ou non de dépression varient significativement en fonction des différentes catégories de durée de sommeil. Pour que ce test soit valide, **la condition de Cochran doit être respectée** (au moins 80 % des effectifs théoriques doivent être supérieurs à 5)**. Si cette condition n'est pas respectée**, **nous utiliserons alors le test exact de Fisher comme alternative**.

➢ Condition de Cochran

Dépression	***État de santé***			
	Healthy	**Moderate**	**Others**	**Unhealthy**
Non	3170,313	4111,998	4,97369	4275,715
Oui	4478,687	5809,002	7,02631	6040,285

Tableau 6 : Tableau des effectifs théoriques des habitudes alimentaires selon la dépression

La fréquence théorique d'une cellule correspond au nombre attendu de sujets dans cette cellule, si les variables étaient indépendantes. Elle se calcule en multipliant les fréquences marginales correspondantes à la ligne et à la colonne de la cellule, puis en divisant le résultat par le nombre total de sujets dans le tableau. **Dans notre cas, la condition de Cochran est respectée, nous pouvons donc faire le test de Khi-deux.**

- **Test de Khi-deux**

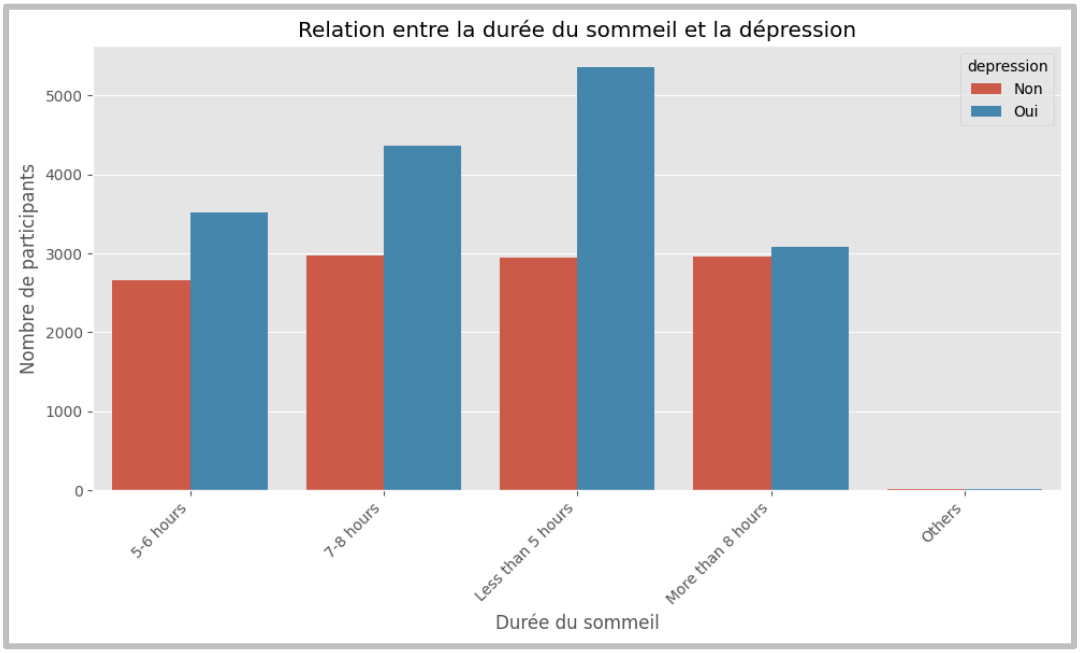
$\mathbf{H_o}$: La dépression est liée aux habitudes alimentaires.

$\mathbf{H_1}$: La dépression du sommeil n'est pas liée aux habitudes alimentaires**.**

La p-value < $2{,}2^{e\text{-}16}$ donc on ne peut rejeter H_o, selon laquelle la dépression est liée aux habitudes alimentaires : **il existe une association significative entre la dépression est liée aux habitudes alimentaires.**

II. LA DURÉE DU SOMMEIL EST-ELLE INDÉPENDANTE DE LA DÉPRESSION ?

1. Test de comparaison graphique



Graphique 5 : Relation entre la durée du sommeil et la dépression

Le **graphique** illustre la relation entre la durée du sommeil et la dépression, en distinguant les participants sans dépression (barres rouges) et avec dépression (barres bleues). On peut observer que :

- **Moins de 5 heures de sommeil** : Cette catégorie présente le plus grand nombre de participants souffrant de dépression, ce qui pourrait indiquer un lien entre une durée de sommeil insuffisante et un risque accru de dépression.
- **5-6 heures de sommeil** : Bien que les effectifs soient légèrement plus faibles que ceux de la catégorie précédente, on note une proportion relativement importante de participants dépressifs.
- **7-8 heures de sommeil** : Cette durée de sommeil est associée à une proportion plus équilibrée entre les participants dépressifs et non dépressifs, ce qui pourrait refléter une association positive avec une meilleure santé mentale.
- **Plus de 8 heures de sommeil** : La proportion de participants non dépressifs est plus élevée que celle des dépressifs, suggérant qu'une durée de sommeil prolongée pourrait être un facteur protecteur contre la dépression.
- **Autres durées** : Ces effectifs sont très faibles, ce qui les rend statistiquement peu significatifs pour l'interprétation.

Dans l'ensemble, on observe que **les personnes ayant une durée de sommeil inférieure à 5 heures présentent la plus forte proportion de dépression (32,819%)**, suivies par celles qui dorment 7-8 heures (26,758%) et 5-6 heures (21,524%). **Les individus qui dorment plus de 8 heures ont une proportion plus faible de dépression (18,843%)**. Les "autres" durées de sommeil, avec des effectifs négligeables, n'ont pas de signification statistique notable.

Ces résultats suggèrent une potentielle association entre des durées de sommeil courtes ou longues et un risque accru ou réduit de dépression.

duree_sommeil	***depression***			
	Non		**Oui**	
	Effectifs	Fréquences	Effectifs	Fréquences
5-6 hours	2665	23,048%	3516	21,524%
7-8 hours	2975	25,729%	4371	26,758%
Less than 5 hours	2948	25,495%	5361	32,819%
More than 8 hours	2966	25,651%	3078	18,843%
Others	9	0,078%	9	0,055%
	11563	1	16335	1

Tableau 7 : Tableau des effectifs de la durée du sommeil selon la dépression

2. Test d'hypothèse

Il s'agit ici d'effectuer un **test du Khi-deux** pour évaluer l'indépendance entre la durée du sommeil et la dépression. L'objectif de ce test est de vérifier si les proportions de participants souffrant ou non de dépression varient significativement en fonction des différentes catégories de durée de sommeil. Pour que ce test soit valide, **la condition de Cochran doit être respectée** (au moins 80 % des effectifs théoriques doivent être supérieurs à 5)**. Si cette condition n'est pas respectée**, **nous utiliserons alors le test exact de Fisher comme alternative**.

➢ Condition de Cochran

duree_sommeil	***depression***	
	Non	**Oui**
5-6 hours	2562	3619
7-8 hours	3045	4301
Less than 5 hours	3444	4865
More than 8 hours	2505	3539
Others	7	11

Tableau 8 : Tableau des effectifs théoriques de la durée du sommeil selon la dépression

La condition de Cochran est respectée, nous pouvons donc faire le test de Khi-deux.

- **<u>Test de Khi-deux</u>**

H_o : La durée du sommeil est liée à la dépression.

H_1 : La durée du sommeil n'est pas liée à la dépression**.**

La p-value < $2,2^{e-16}$ donc on ne peut rejeter H_o, selon laquelle la durée du sommeil est indépendante de la dépression : **il existe une association significative entre la durée du sommeil et la dépression.**

CONCLUSION ET RECOMMANDATIONS

Cette étude visait à analyser les facteurs influençant la dépression chez les étudiants en se basant sur des méthodes statistiques. À travers une approche rigoureuse, plusieurs étapes ont été réalisées, notamment :

- L'analyse descriptive des données relatives à la dépression et à des facteurs associés tels que la pression académique, le stress financier, et les habitudes de sommeil, qui a permis de dégager des tendances clés et de mieux comprendre la répartition de la dépression chez les étudiants ;
- L'estimation des proportions et des intervalles de confiance, qui a permis d'évaluer la prévalence de certains facteurs de risque comme les pensées suicidaires, ainsi que la moyenne et la médiane des heures de travail ou d'études pour les étudiants souffrant de dépression ;
- Le test d'hypothèses, qui a permis d'analyser la satisfaction académique et professionnelle en fonction de la dépression et de vérifier l'indépendance entre la dépression et des variables comme les habitudes alimentaires ou la durée du sommeil.

Les résultats obtenus ont permis de révéler plusieurs liens significatifs entre la dépression et des facteurs tels que la **durée de sommeil**, le **stress financier**, et la **satisfaction des études et du travail**. Ces résultats suggèrent que **la dépression chez les étudiants est influencée par une combinaison de facteurs académiques, sociaux et de bien-être**.

En conclusion, cette étude a démontré l'importance d'une analyse statistique approfondie pour comprendre les facteurs contribuant à la dépression chez les étudiants. Les **recommandations** proposées, telles que l'**amélioration des stratégies de gestion du stress et de soutien psychologique**, pourraient être exploitées pour promouvoir un environnement académique plus sain et améliorer le bien-être des étudiants.

ANNEXES

1. Bibliographie

Akposso, D. (2022). *Statistiques Inférentielles* : *Support de cours*. Institut Supérieur de Statistique d'Économétrie et de Data Science.

2. Source du code Python

http://bit.ly/4hhwYac