MINISTERE DE L'ENSEIGNEMENT SUPERIEURET DE RECHERCHE

REPUBLIQUE DE COTE D'IVOIRE

Institut Supérieur de Statistique d'Econométrie et de Data Science

STATISTIQUE INFERENTIELLE

ANNEE ACADEMIQUE 2024 - 2025

Etudiant

ANGORAN

KOUAME GILLES

Enseignant – Encadreur

AKPOSSO DIDIER

MARTIAL

Table des matières

INTRODUCTION GENERALE

Contexte et justification de l'étude

La santé mentale des étudiants est une question de plus en plus discutée, car elle touche directement le bien-être, la réussite académique et la trajectoire personnelle. Parmi les difficultés rencontrées, la dépression occupe une place centrale : elle peut altérer la motivation, la concentration et l'engagement dans les études, avec des conséquences parfois graves comme l'isolement ou les idées suicidaires. Dans ce contexte, il devient pertinent d'examiner les facteurs associés à la dépression afin d'orienter plus efficacement la prévention et les actions de soutien.

Cette étude s'appuie sur une base de données décrivant des étudiants à travers des informations démographiques (âge, sexe, ville), académiques (pression académique, moyenne, satisfaction des études), et liées au mode de vie (durée du sommeil, habitudes alimentaires), ainsi que des éléments psychosociaux (stress financier, antécédents familiaux). La variable d'intérêt est la présence de dépression (codée 1/0), et une attention particulière est portée aux pensées suicidaires (Oui/Non).

L'objectif est d'estimer certains indicateurs clés (notamment un intervalle de confiance pour la proportion d'étudiants ayant eu des pensées suicidaires) et de tester, par des méthodes statistiques inférentielles, l'existence d'associations significatives entre la dépression et plusieurs facteurs explicatifs. Les analyses seront réalisées sous R, en distinguant clairement association statistique et interprétation causale.

Principaux résultats attendus

L'objectif de cette étude est d'identifier les **facteurs associés** à la dépression chez les étudiants, à partir de la variable binaire *depression* (1 = oui, 0 = non). Les analyses viseront à :

- **Estimer** la proportion d'étudiants ayant déjà eu des pensées suicidaires et fournir un **intervalle de confiance à 95%**.
- **Quantifier** les différences de niveau de stress financier et d'heures de travail/études entre les groupes (dépressifs vs non dépressifs) à l'aide de mesures de tendance centrale (moyenne, médiane) et d'indicateurs adaptés.
- **Tester** si la satisfaction des études diffère significativement entre étudiants dépressifs et non dépressifs..
- **Évaluer** l'existence d'une association entre la dépression et certaines variables qualitatives (habitudes alimentaires, durée du sommeil, antécédents familiaux de maladies mentales).
- Examiner, si les données le permettent, la relation entre **satisfaction au travail** et **diplôme suivi**, en tenant compte des effectifs réellement concernés.

Les résultats attendus sont donc des **estimations**, des **tests statistiques** et des **conclusions d'association**, qui pourront éclairer des recommandations pratiques, tout en rappelant les limites liées à la nature observationnelle des données.

Méthodologie

L'étude est réalisée sous **R** et suit une démarche structurée en quatre étapes :

1. **Préparation et contrôle qualité des données**
   Nettoyage des valeurs manquantes, harmonisation des codages (Oui/Non, 0/1), vérification des types de variables, détection d'incohérences et de variables quasi constantes. Une attention particulière est portée aux variables liées au travail afin de distinguer les valeurs réellement observées des valeurs "non concerné".

2. **Analyses descriptives préliminaires**

- **Analyse univariée** : description des variables quantitatives (moyenne, médiane, dispersion) et qualitatives (effectifs, pourcentages).
- **Analyse bivariée descriptive** : comparaison exploratoire des distributions selon le statut de dépression, afin de préparer les tests inférentiels.

3. **Analyses inférentielles (seuil $\alpha = 5\%$)**

- **Intervalle de confiance** pour la proportion d'étudiants ayant eu des pensées suicidaires (approche binomiale).
- **Comparaisons entre groupes** (dépressifs vs non dépressifs) : tests adaptés à la nature des variables (test t de Welch ou Wilcoxon pour des scores, selon la distribution ou autre).
- **Associations entre variables qualitatives** : tests du **chi-deux** (ou alternatives si les effectifs attendus sont insuffisants).
- **Comparaison selon diplôme** pour la satisfaction au travail : analyses réalisées uniquement si les effectifs exploitables sont suffisants, avec un commentaire méthodologique en cas de limitation.

4. **Visualisation et restitution**
   Export d'un jeu de données propre et production d'un **tableau de bord Power BI** synthétisant les résultats (indicateurs clés, comparaisons, distributions). Rédaction d'un rapport final présentant les résultats, leur interprétation et des recommandations, en distinguant clairement association statistique et causalité.

Le dictionnaire de données

Un dictionnaire de données est un document qui fournit une description détaillée de toutes les variables utilisées dans une analyse statistique. Il décrit les propriétés et les caractéristiques de chaque variable , ainsi que leur signification selon le contexte.

Groupe de variables	Nom de la variable	Nature	Description	Modalités
Identifiant	id	Numérique	Identifiant unique de chaque étudiant	Valeurs uniques
Démographique	sexe	Catégorielle	Sexe de l'étudiant	Masculin / Féminin
Démographique	age	Numérique	Âge de l'étudiant	Nombre entier (années)
Démographique	ville	Catégorielle	Ville de résidence de l'étudiant	Noms de villes
Scolaire	profession	Catégorielle	Statut professionnel de l'étudiant	Étudiant à temps plein / Autre activité
Scolaire	pression_academique	Ordinale	Niveau de pression académique ressenti	Échelle (faible à élevée)
Scolaire	pression_liee_au_travail	Ordinale	Niveau de pression liée au travail	Échelle (faible à élevée)
Scolaire	moyenne_notes	Numérique	Moyenne générale des notes de l'étudiant	Note sur une échelle donnée
Scolaire	satisfaction_etudes	Ordinale	Niveau de satisfaction par rapport aux études	Échelle (faible à élevée)
Scolaire	satisfaction_travail	Ordinale	Niveau de satisfaction par rapport au travail	Échelle (faible à élevée)
Mode de vie	duree_sommeil	Catégorielle	Durée moyenne du sommeil par nuit	Moins de 5h / 5-6h / 7-8h / Plus de 8h
Mode de vie	habitudes_alimentaires	Catégorielle	Type d'habitudes alimentaires	Saines / Modérées
Scolaire	diplome_suivi	Catégorielle	Diplôme en cours ou obtenu	BSc / M.Tech / Autre
Santé mentale	pensees_suicidaires	Binaire	Existence de pensées suicidaires	Oui / Non
	nombre_heure_travail_etude	Numérique	Nombre d'heures consacrées aux études ou au travail	Nombre entier (heures)
Financier	stress_financier	Ordinale	Niveau de stress financier	Échelle (faible à élevée)
Santé mentale	antecedents_familiaux_maladies_mentales	Binaire	Antécédents familiaux de maladies mentales	Oui / Non
Santé mentale	depression	Binaire	Présence de symptômes de dépression	1 = Oui / 0 = Non

Voici ainsi présenté chaque variable ces caractéristiques et c'est spécificités

I.ETUDE DE LA BASE DE DONNEES

Cette étape sert à faire une chose simple mais non négociable : **vérifier que la base est exploitable** avant de sortir des tests et des p-values. Concrètement, on décrit d'abord la **structure** du fichier (dimensions, liste des variables, types, codages), puis on réalise un contrôle qualité orienté analyse : **valeurs manquantes**, **déséquilibres de modalités**, **valeurs extrêmes**, et incohérences possibles. L'objectif est d'éviter deux erreurs fréquentes : tester des variables mal codées (ce qui donne des "résultats" trompeurs) et ignorer des limitations évidentes du jeu de données.

1. Présentation du jeu de données

Nous avons pris le soin de renominer quelques variables, des encodages et des changements de nature de quelques variables afin de favoriser une manipulation plus simple et une meilleure analyse.

	id <fctr>	sexe <fctr>	age <dbl>	ville <fctr>	profession <fctr>	pression_academique <dbl>	pression_liee_au_travail <dbl>	moyenne_notes <dbl>
1	2	Male	33	Visakhapatnam	Student	5	0	8.97
2	8	Female	24	Bangalore	Student	2	0	5.90
3	26	Male	31	Srinagar	Student	3	0	7.03
4	30	Female	28	Varanasi	Student	3	0	5.59

satisfaction_etudes <dbl>	satisfaction_travail <dbl>	duree_sommeil <fctr>	habitudes_alimentaires <fctr>	diplome_suivi <fctr>	pensees_suicidaire <fctr>
2	0	5-6 hours	Healthy	B.Pharm	Yes
5	0	5-6 hours	Moderate	BSc	No
5	0	Less than 5 hours	Healthy	BA	No
2	0	7-8 hours	Moderate	BCA	Yes

pensees_suicidaire <fctr>	nombre_heure_travail_etude <dbl>	stress_financier <dbl>	antecedants_familiaux_maladie_mentale <fctr>	depression <fctr>
Yes	3	1	No	oui
No	3	2	Yes	non
No	9	1	Yes	non
Yes	4	5	Yes	oui

Le jeu de données contient **27 901 observations** décrites par **18 variables**. Il regroupe des informations démographiques et de contexte (id, sexe, âge, ville, profession, diplome_suivi), des indicateurs académiques (pression académique, moyenne notes, satisfaction études), des facteurs de mode de vie (durée_sommeil, habitudes_alimentaires, nombre_heure_travail_etude) ainsi que des variables directement liées à la santé mentale (pensees_suicidaire, antecedants_familiaux_maladie_mentale, depression). Deux variables portent sur la dimension "travail" (pression_liee_au_travail, satisfaction_travail), mais leur interprétation devra être traitée avec prudence, car elles sont très souvent nulles dans la base.

D'un point de vue statistique, les variables se répartissent en deux familles. D'une part, des **variables quantitatives** (mesures ou scores) : âge, pression_academique, pression_liee_au_travail, moyenne_notes, satisfaction_etudes, satisfaction_travail, nombre_heure_travail_etude, stress_financier. D'autre part, des **variables qualitatives** : sexe, ville, profession, duree_sommeil, habitudes_alimentaires, diplome_suivi, pensees_suicidaire, antecedants_familiaux_maladie_mentale, dépression. Il est toutefois important de noter que plusieurs variables quantitatives sont en réalité des **scores discrets et ordonnés** (par exemple la pression, la satisfaction ou le stress), ce qui influencera le choix des tests (tests robustes/non paramétriques en complément des tests paramétriques).

L'étape suivante consistera à approfondir le diagnostic de qualité et de structure des données : repérer d'éventuelles **valeurs extrêmes** sur les variables numériques, confirmer

et documenter la **répartition des valeurs manquantes**, puis produire des indicateurs de **tendance centrale et de dispersion** (moyenne, médiane, écart-type, IQR) afin de dégager la tendance générale de chaque variable et préparer l'analyse univariée et bivariée.

2. Traitement des doublons

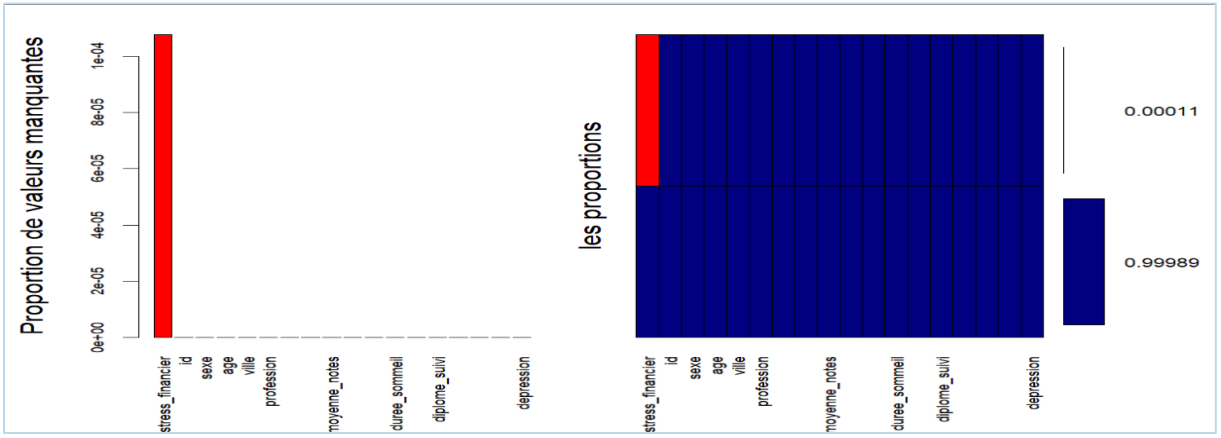
```
#[1] 0
```

La vérification d'unicité de la variable id confirme l'absence de doublons dans la base : aucun identifiant n'est répété.

3. Traitement des valeurs manquantes

Cette étape vise à identifier les observations incomplètes, à mesurer précisément la proportion de valeurs manquantes pour chaque variable, puis à retenir une stratégie de traitement cohérente avec l'objectif de l'analyse. L'enjeu n'est pas de remplacer systématiquement les valeurs absentes, mais de préserver la fiabilité des résultats en évitant à la fois la perte inutile d'information et l'introduction de biais par un traitement inadapté.

3.1 Vérification des valeurs manquantes dans la base de données



Nombre de valeurs manquantes

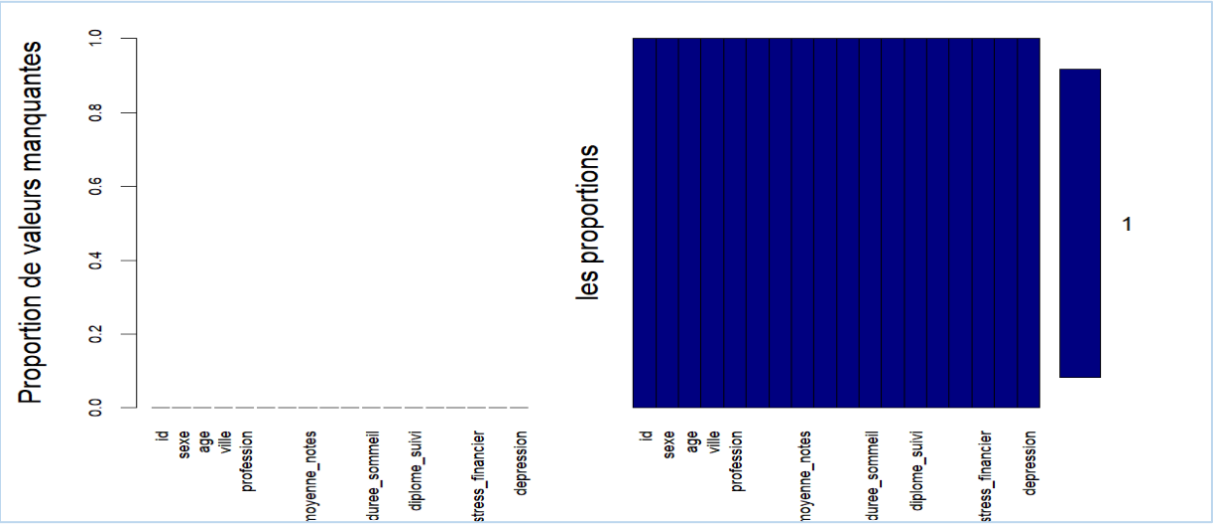
```
## [1] 3
```

Le graphique montre que les valeurs manquantes sont **quasi inexistantes** dans la base : elles concernent uniquement la variable **stress_financier**, avec **3 valeurs manquantes** sur **27 901** (≈ **0,01%**). Toutes les autres variables sont complètes, ce qui confirme une très bonne qualité de données.

Comme stress_financier est un **score ordinal (1–5)** potentiellement important pour l'analyse, nous privilégions une **imputation par la médiane**. Ce choix permet de conserver l'intégralité de l'échantillon, tout en évitant d'introduire un biais artificiel : la médiane est robuste et moins sensible aux valeurs extrêmes qu'une moyenne, ce qui la rend adaptée dans ce contexte.

3.2. Traitement des données manquantes

Nous allons remplacer les valeurs manquantes par la médiane




```
## [1] 0
```

Sur le visuel nous pouvons remarquer que le données manquantes ont été traité, nous avons 0 données manquantes.

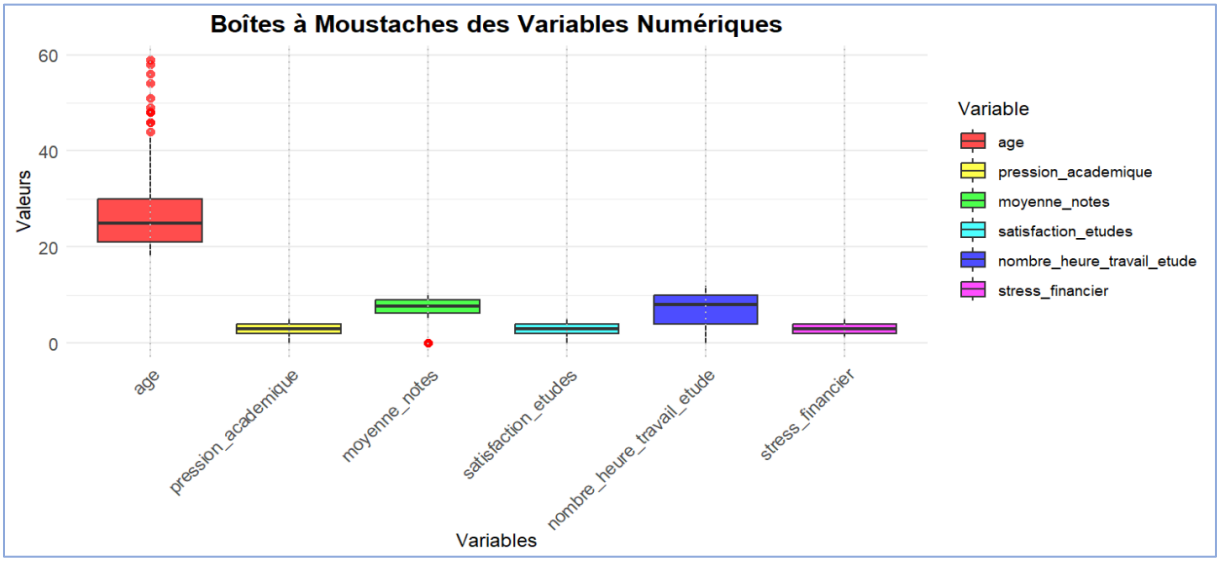
4. Taitement des valeurs abbérantes

À cette étape, l'objectif est de repérer d'éventuelles valeurs extrêmes ou incohérentes dans les **variables numériques** susceptibles de perturber l'analyse descriptive et certains tests. Dans notre base, les variables concernées sont notamment age, moyenne_notes, nombre_heure_travail_etude et, dans une moindre mesure, les scores pression_academique, satisfaction_etudes et stress_financier. Même si plusieurs de ces variables sont **bornées** (par exemple 0–5 ou 1–5), il reste utile de vérifier l'existence de valeurs atypiques, d'autant plus que ces valeurs peuvent affecter les moyennes et la dispersion.
La détection s'appuiera d'abord sur une approche graphique, notamment la **boîte à moustaches**, qui permet de visualiser rapidement la distribution et d'identifier les points atypiques. Si des valeurs extrêmes apparaissent, elles seront ensuite examinées au cas par cas :

- si elles sont **plausibles** (ex. âge élevé mais possible, heures/jour élevées mais compatibles avec l'échelle), elles seront conservées ;
- si elles traduisent une **incohérence** ou une erreur (ex. âge hors intervalle réaliste, notes hors [0–10], heures hors [0–12]), elles seront corrigées ou exclues avec justification.

L'idée n'est pas de "nettoyer" pour rendre les résultats plus jolis, mais de garantir que les analyses inférentielles reposent sur des données cohérentes et interprétables.

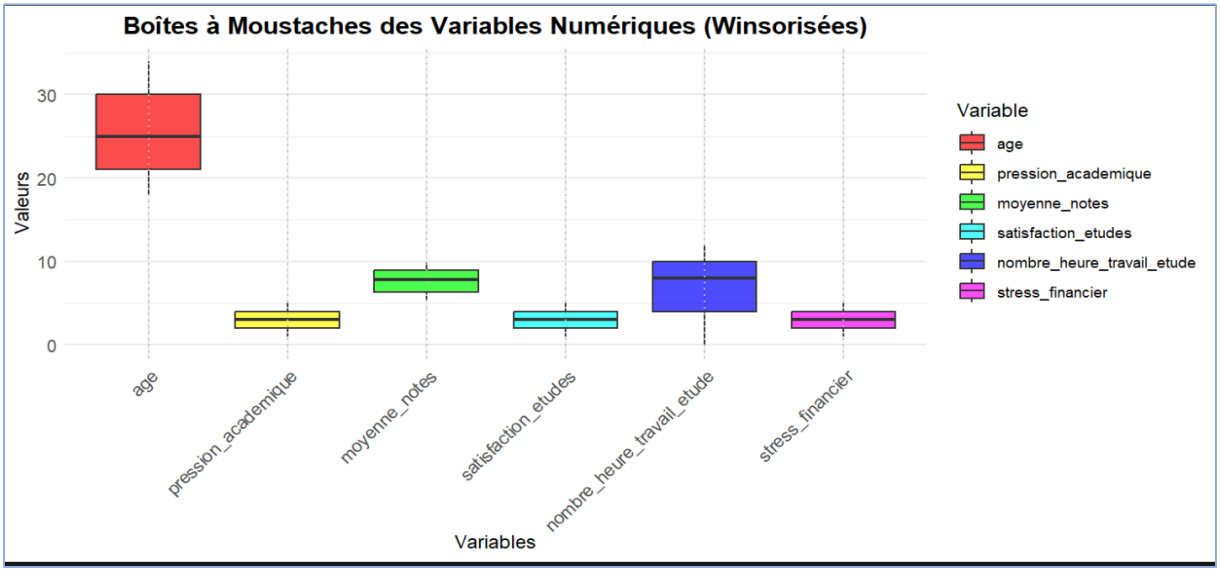
4.1. Affichage de boite à moustache



Ce visuel nous présente effectivement des valeurs extremes qui neccessite d'être traitées et pour ce faire nous allons utiliser la technique de winsorisation

4.2. Winsorisation des valeurs extremes

La winorization est une méthode statistique qui permet de traiter les valeurs extrêmes dans un jeu de données. Elle consiste à remplacer les valeurs extrêmes par des valeurs plus proches qui ne sont pas considérées comme extrêmes, les valeurs maximales et minimales des boites à moustaches.



Nous pouvons remarquer après la winsorisation que toutes les valeurs extremes ont été traitées. Nous pouvons à présent passer à l'analyse univarié, l'analyse univariée des variables permet de dégager des informations cruciales pour comprendre les caractéristiques de chaque indicateur.

II. QUELQUES ETUDES PRELIMINAIRES

1- ANALYSE UNIVARIEES

.

1.1 . Tableau statistique simple des variables quantitative

Variable <chr>	moyenne <dbl>	mediane <dbl>	ecart_type <dbl>	minimum <dbl>	q25 <dbl>	q75 <dbl>	maximum <dbl>	skewness <dbl>	interpskew <chr>	kurtosis <dbl>	interpkurt <chr>
age	25.82	25.00	4.91	18	21.00	30.00	59	0.13	distribution étalée à droite	2.15	platikurtique
pression_academique	3.14	3.00	1.38	0	2.00	4.00	5	-0.14	distribution étalée à gauche	1.84	platikurtique
moyenne_notes	7.66	7.77	1.47	0	6.29	8.92	10	-0.11	distribution étalée à gauche	1.98	platikurtique
satisfaction_etudes	2.94	3.00	1.36	0	2.00	4.00	5	0.01	distribution étalée à droite	1.78	platikurtique
nombre_heure_travail_etude	7.16	8.00	3.71	0	4.00	10.00	12	-0.45	distribution étalée à gauche	2.00	platikurtique

Le tableau ci-dessous présente les statistiques descriptives des différentes variables, calculées à l'aide du logiciel R

1.2. Interprétation des paramètres du tableau statistique simple

1.2.1. Paramètre de tendance centrale

Moyenne
L'âge moyen est de 25,82 ans, ce qui confirme une population majoritairement jeune. La pression académique est en moyenne de 3,14 (sur 0–5), indiquant un niveau globalement modéré. La moyenne des notes est de 7,66/10, ce qui suggère des performances scolaires globalement bonnes. La satisfaction des études est en moyenne de 2,94 (sur 0–5), traduisant une satisfaction intermédiaire. Enfin, les étudiants consacrent en moyenne 7,16 heures/jour au travail/études, ce qui reflète une charge quotidienne importante.

Minimum
Les minimums sont globalement cohérents (âge min = 18). En revanche, certaines variables affichent un minimum à 0 (notes, pression, satisfaction, heures). Pour nombre_heure_travail_etude, 0 peut être interprété comme "aucune heure" (possible). Pour moyenne_notes et les scores, un 0 mérite d'être vérifié : valeur réelle ou codage particulier selon la source du dataset.

Maximum
Les maximums restent compatibles avec les échelles : âge max 59, pression académique 5, satisfaction des études 5, notes 10, heures 12. Cela montre l'existence d'étudiants en situation "extrême" sur certains axes (pression élevée, forte charge horaire), ce qui peut jouer dans les comparaisons avec la dépression.

Médiane
Les médianes confirment le profil majoritaire : âge médian 25, pression 3, notes 7,77, satisfaction 3, heures 8. Le fait que moyenne et médiane soient proches pour la plupart des variables indique des distributions plutôt équilibrées (à confirmer avec la forme).

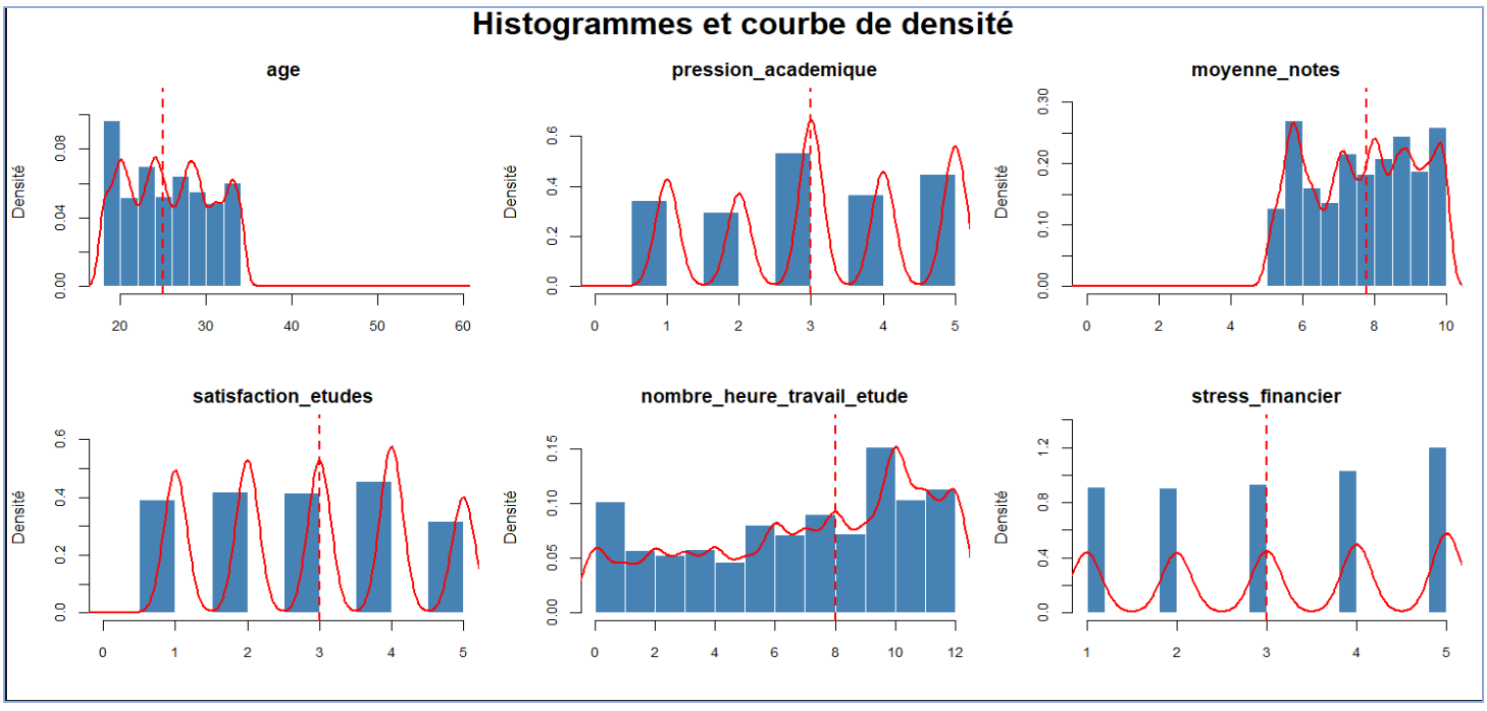
1.2.2. Paramètre de dispertion

Âge : dispersion modérée (écart-type = 4,91) ; 50% des étudiants ont un âge entre 21 et 30 ans (IQR = 9).
Pression académique : variabilité notable (écart-type = 1,38) avec la moitié des valeurs entre 2 et 4, donc des ressentis assez hétérogènes.
Moyenne des notes : dispersion modérée (écart-type = 1,47) ; 50% des notes entre 6,29 et 8,92, ce qui traduit des niveaux scolaires variés mais globalement élevés.
Satisfaction des études : dispersion comparable (écart-type = 1,36) ; 50% des valeurs entre 2 et 4, donc perceptions partagées.
Heures de travail/études : c'est la variable la plus dispersée (écart-type = 3,71) avec un IQR large (4 à 10 heures). Elle différencie fortement les profils et sera probablement informative pour l'analyse bivariée.

1.2.3. Paramètre de forme

Âge : asymétrie très faible (**skewness = 0,13**), distribution légèrement étalée à droite ; kurtosis **2,15** (platikurtique) → distribution assez "plate", sans pic excessif.
Pression académique : **skewness = -0,14** (légère asymétrie à gauche), kurtosis **1,84** → distribution plutôt étalée, sans concentration extrême.
Moyenne des notes : **skewness = -0,11** (légère asymétrie à gauche), kurtosis **1,98** → tendance à des notes relativement élevées pour une partie importante des étudiants.
Satisfaction des études : **skewness = 0,01** (quasi symétrique), kurtosis **1,78** → répartition assez régulière autour du niveau moyen.
Heures de travail/études : **skewness = -0,45** (asymétrie à gauche), kurtosis **2,00** → beaucoup d'étudiants sont plutôt vers les valeurs élevées, avec une queue vers les faibles volumes (quelques profils à très peu d'heures).

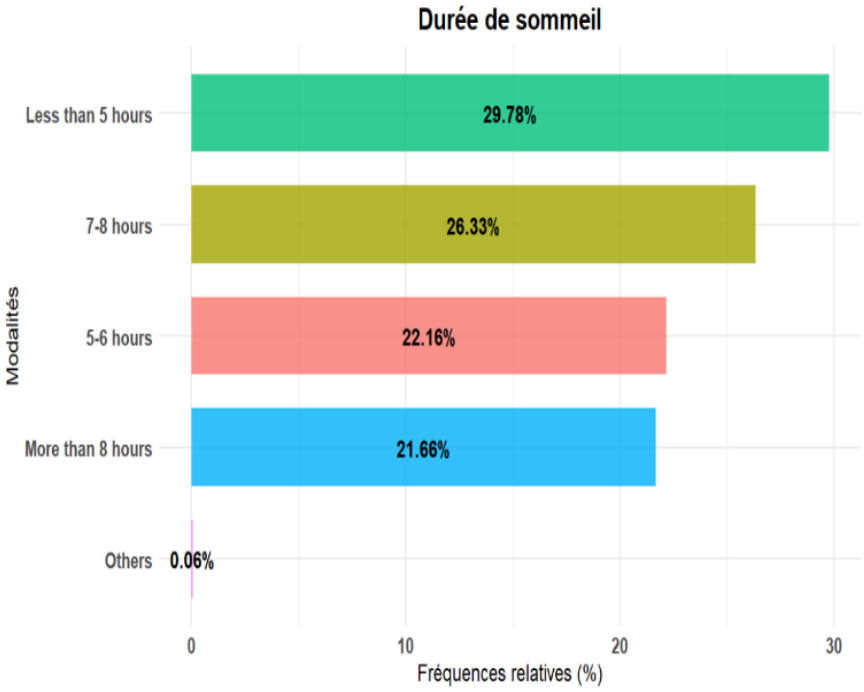
Histogramme de nos variables quantitatives pour confirmer nos oresultats numériques



1.3 Analyse des variables qualitatives

1.3.1 Représentation graphique de la variable durée_sommeil

- Ci-dessous, nous avons une description visuelle et écrite de la variable Durée de sommeil



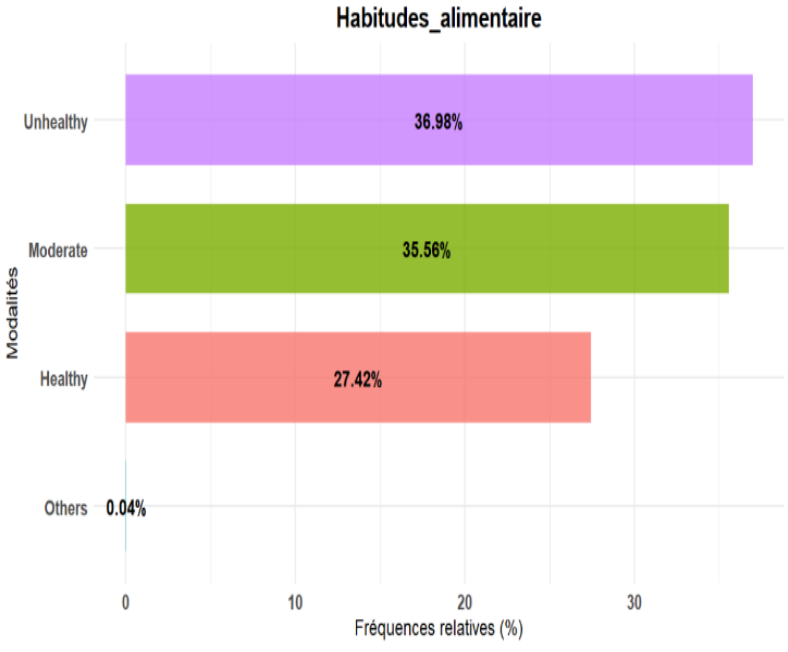
La durée de sommeil inférieure à **5 heures** est la plus fréquente **(29,78 %),** traduisant une proportion importante d'individus en situation de privation de sommeil.

Les durées intermédiaires **(5–6 heures et 7–8 heures**) regroupent près de la moitié de l'échantillon, révélant des profils hétérogènes.

La catégorie "plus de 8 heures" reste non négligeable **(21,66 %),** tandis que les autres modalités sont marginales et sans impact analytique.

1.3.2 Représentation graphique de la variable Habitudes_alimentaires

- Ci-dessous, nous avons une description visuelle et écrite de la variable *Habitudes_alimentaires*



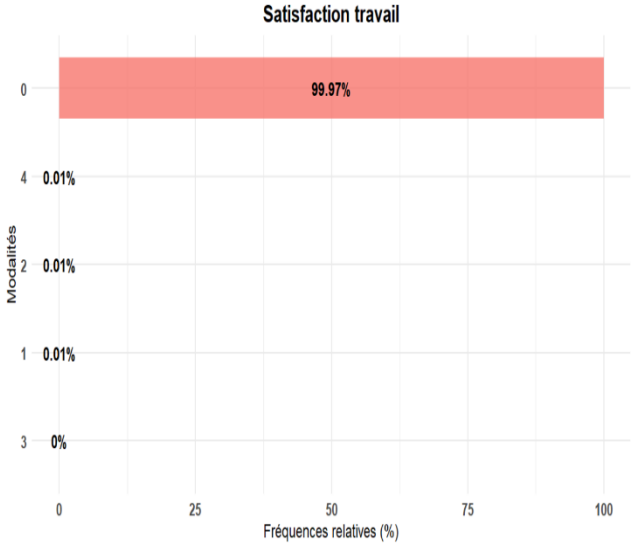
Les habitudes alimentaires sont majoritairement défavorables ou intermédiaires, avec **36,98 %** d'individus ayant une alimentation malsaine et **35,56 %** une alimentation modérée.

L'alimentation saine reste minoritaire **(27,42 %),** traduisant un déséquilibre global au profit de comportements nutritionnels peu favorables.

La catégorie "Others" est marginale et n'a aucun poids analytique.

1.3.3 Représentation graphique de la variable Satisfaction_travail

- Ci-dessous, nous avons une brève description visuelle et écrite de la variable *Satisfaction_travail*

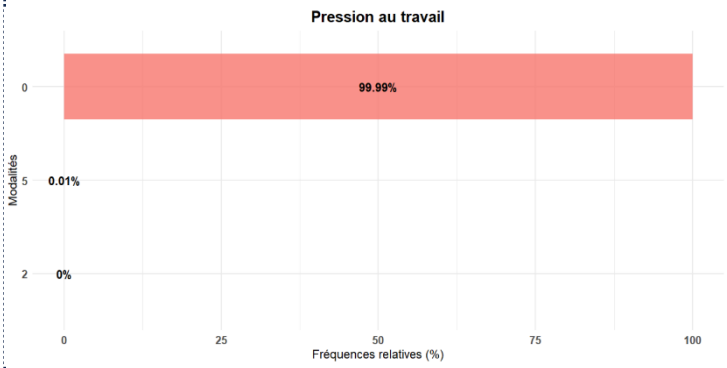


La quasi-totalité des individus se situe au niveau de satisfaction le plus faible (niveau 0), représentant **99,97 %** de l'échantillon.

Les niveaux de satisfaction plus élevés sont pratiquement inexistants, chacun représentant environ **0,01 %** ou moins.
Cette distribution traduit une insatisfaction massive et généralisée au travail, sans réelle variabilité entre les individus.

1.3.4 Représentation graphique de la Pression au travail

- Ci-dessous, nous avons une brève description visuelle et écrite de la variable *Pression au travail*



La variable "pression au travail" est quasi entièrement concentrée sur le niveau **0 (99,99 %),** les autres niveaux étant statistiquement inexistants.

Cette distribution extrême indique une absence totale de variabilité, rendant toute interprétation comportementale ou sanitaire non exploitable en l'état.

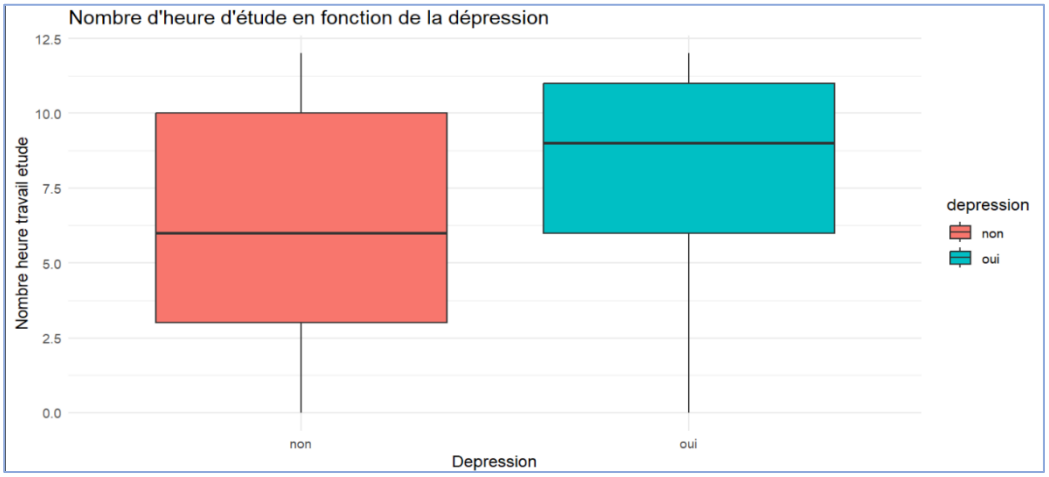
1.4 Synthèse de l' analyse Univarié

L'analyse univariée met en évidence une **variabilité marquée des comportements sanitaires**, notamment en matière de sommeil et d'alimentation, suggérant des profils de santé contrastés au sein de l'échantillon.
À l'inverse, les variables liées au travail présentent une **absence quasi totale de dispersion**, limitant fortement leur portée interprétative en l'état.

L'analyse bivariée vise à explorer les **relations potentielles entre les variables présentant une variabilité effective**.
Elle permet de formuler des **soupçons analytiques** quant à l'existence de liens entre comportements, soupçons qui seront ensuite **testés et validés — ou infirmés — par nos tests**.

2. ANALYSE BIVARIEE

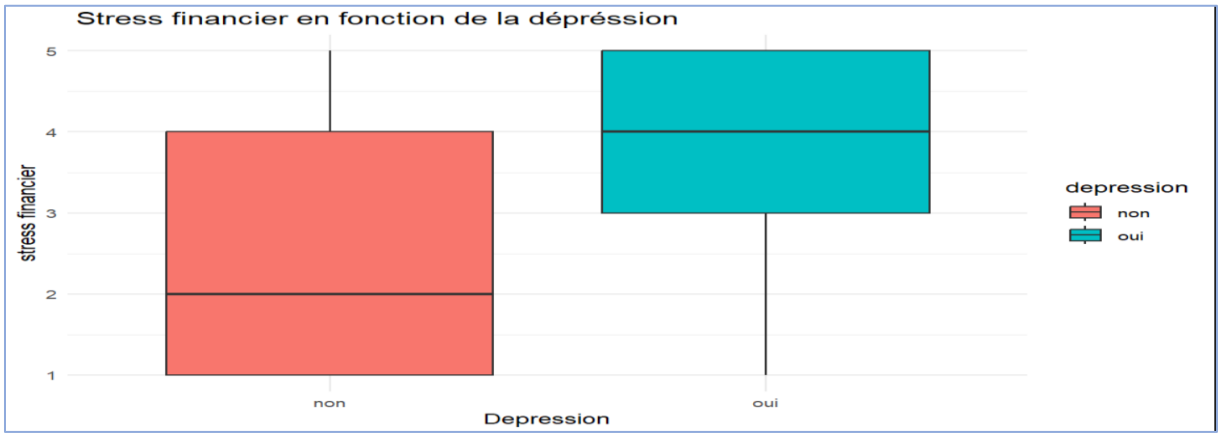
2.1 Analyse bivarié entre le nombre d'heure d'étude en fonction de la dépression



À la lecture du boxplot, on observe que les étudiants **dépressifs** présentent globalement un **nombre d'heures de travail/études plus élevé** que ceux qui ne le sont pas : la **médiane** et l'ensemble de la distribution sont décalés vers le haut dans le groupe *oui*. Cela suggère une **association** entre la dépression et une charge quotidienne plus importante.

Cependant, ce graphique ne permet pas de conclure à un effet causal. Il motive plutôt l'hypothèse d'une **différence de niveau** entre les deux groupes, hypothèse que nous confirmerons (ou infirmerons) à l'aide d'un **test de comparaison approprié** (par exemple Wilcoxon ou Welch selon les conditions).

2.2 Analyse bivarié entre Stress financier et depression

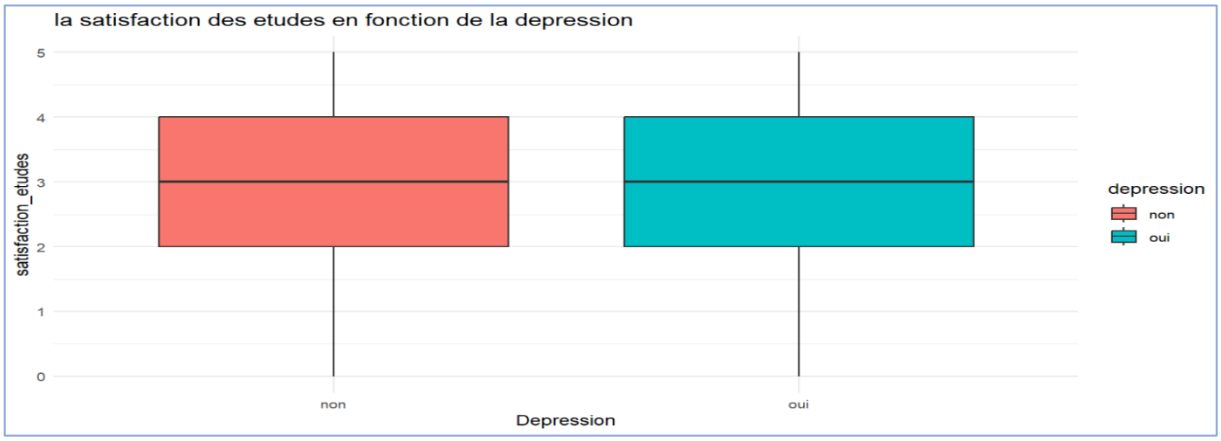


Au vu du boxplot, on observe un **écart marqué** entre les deux groupes. Les étudiants **dépressifs (oui)** présentent un **stress financier nettement plus élevé** que ceux qui ne le sont pas : la **médiane** du groupe *oui* se situe autour de **4**, tandis que celle du groupe *non* est proche de **2**. De plus, la majorité des valeurs chez les dépressifs se concentre entre **3 et 5**, alors que les non dépressifs sont davantage répartis vers les niveaux faibles (1–2).

Ces résultats nous amènent donc à **soupçonner une association** entre la dépression et un

niveau de stress financier plus important. Toutefois, ce graphique ne permet pas de conclure à un effet causal ; il motive plutôt l'hypothèse d'une **différence de distribution** entre les deux groupes, que nous confirmerons (ou infirmerons) à l'aide d'un **test de comparaison approprié** (idéalement Wilcoxon, car il s'agit d'un score ordinal)

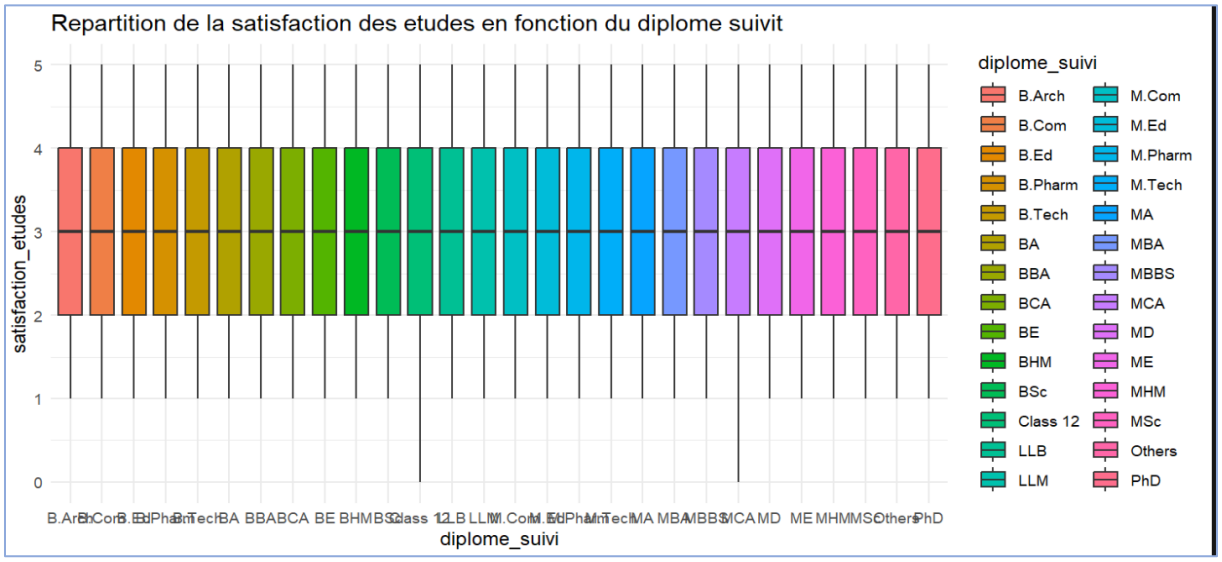
2.3 Analyse bivarié entre la satisfaction des études en fonction de la dépression



Au vu du boxplot, on remarque que la **satisfaction des études** est globalement **très similaire** chez les étudiants dépressifs et non dépressifs. Les deux groupes présentent une **médiane identique**, autour de **3**, et des distributions qui se recouvrent fortement : la moitié des observations se situe approximativement entre **2 et 4** dans les deux cas, avec des valeurs extrêmes comparables.

Autrement dit, ce graphique ne met pas en évidence une différence visuelle marquée de satisfaction des études entre les deux groupes. Cela nous amène donc à soupçonner que la satisfaction des études pourrait être **faiblement liée** à la dépression dans ce jeu de données, voire **indépendante**. Toutefois, cette impression doit être confirmée par un **test de comparaison** (idéalement un test de Wilcoxon, car il s'agit d'un score ordinal), afin de vérifier si une différence existe malgré le fort chevauchement observé.

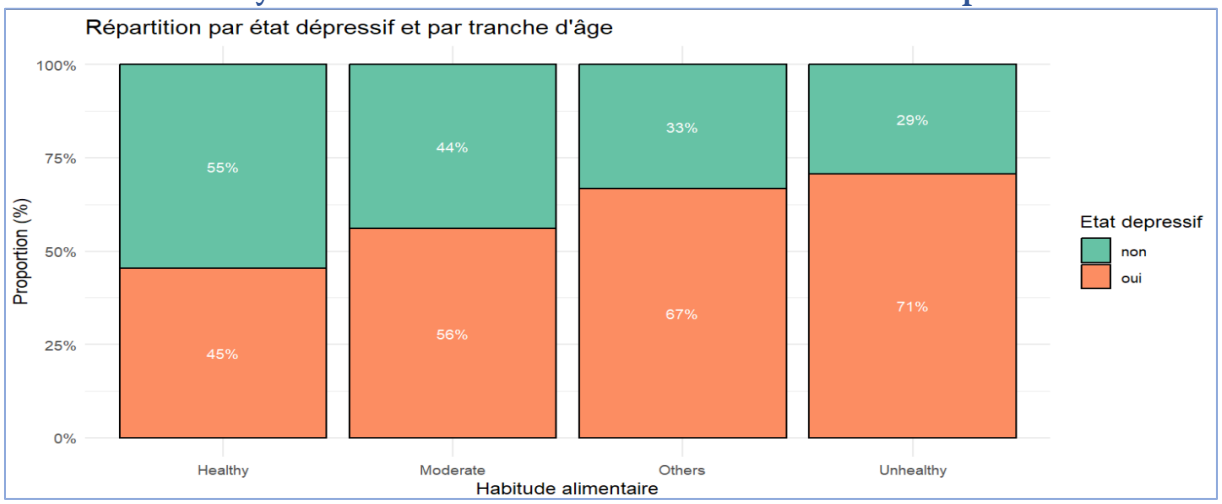
2.4 Analyse bivarié entre la satisfaction des études en fonction du diplôme suivit



Au vu du graphique, on remarque que la satisfaction des études varie très peu d'un diplôme à l'autre. La plupart des boxplots sont quasiment superposés : la médiane se situe presque toujours autour de 3, et l'intervalle interquartile est généralement compris entre 2 et 4. Autrement dit, quel que soit le diplôme suivi, la satisfaction reste globalement moyenne et relativement stable.
On observe bien quelques moustaches plus longues ou quelques valeurs plus basses dans certains diplômes, mais rien de suffisamment structuré visuellement pour affirmer qu'un diplôme se distingue nettement des autres. Cela nous amène à soupçonner que le diplôme suivi n'explique pas fortement la satisfaction des études dans ce jeu de données.

Cependant, comme il y a beaucoup de catégories (plusieurs diplômes, avec probablement des effectifs très différents), cette impression doit être confirmée par un test statistique adapté : par exemple un test de Kruskal-Wallis (score ordinal) ou une ANOVA si l'on accepte l'hypothèse de normalité/variance, puis éventuellement des comparaisons post-hoc si le test global est significatif.

2.5 Analyse bivarié entre les habitudes alimentaires et la depression



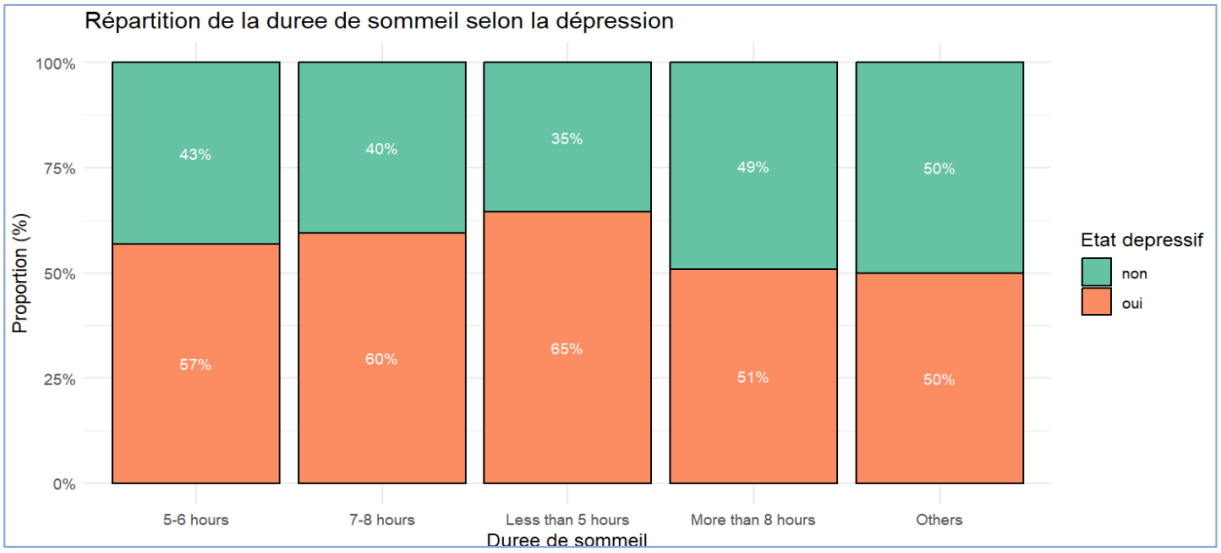
Au vu du graphique (barres empilées en pourcentage), on observe une **tendance nette**

entre les habitudes alimentaires et l'état dépressif.

- Pour les étudiants ayant des habitudes **Healthy**, la majorité est **non dépressive** (**55% non** contre **45% oui**).
- Dès qu'on passe à **Moderate**, la proportion s'inverse : les **dépressifs deviennent majoritaires** (**56% oui** contre **44% non**).
- La catégorie **Unhealthy** montre l'écart le plus marqué : environ **71% oui** contre **29% non**, ce qui suggère une association forte entre habitudes alimentaires défavorables et dépression.
- La modalité **Others** présente aussi une proportion élevée de dépression (**67% oui**), mais cette catégorie est généralement **très faible en effectif** dans ton dataset, donc elle doit être interprétée avec prudence.

En résumé, plus les habitudes alimentaires se dégradent (de Healthy vers Unhealthy), plus la proportion d'étudiants dépressifs augmente. Cela nous amène à soupçonner une **dépendance** entre habitudes_alimentaires et depression, hypothèse que nous vérifierons formellement avec un **test du chi-deux d'indépendance** en cas de normalité (en regroupant ou en excluant la modalité Others si les effectifs attendus sont trop faibles).

2.6 Analyse bivarié entre la durée de sommeil et la depression



Au vu du graphique, on observe une relation assez claire entre la **durée du sommeil** et l'état dépressif : les proportions de dépression ne sont pas les mêmes selon les catégories de sommeil.

- Chez les étudiants dormant **moins de 5 heures**, la dépression est la plus fréquente : **65%** sont dépressifs contre **35%** non dépressifs.
- Pour les durées **5–6 heures** et **7–8 heures**, la dépression reste majoritaire, avec respectivement **57%** et **60%** d'étudiants dépressifs.
- En revanche, lorsque la durée de sommeil dépasse **8 heures**, la tendance s'atténue fortement : on obtient presque un équilibre, avec **51%** de dépressifs contre **49%** de non dépressifs.
- La catégorie **Others** est proche de l'équilibre (**50% / 50%**), mais elle est généralement peu représentée, donc à interpréter avec prudence.

Globalement, plus la durée de sommeil est faible, plus la proportion d'étudiants

dépressifs augmente, ce qui nous amène à soupçonner une **dépendance** entre duree_sommeil et depression. Cette hypothèse devra être confirmée par un **test du chi-deux d'indépendance**, en vérifiant au préalable les effectifs attendus (et en traitant la modalité Others si nécessaire).

2.7 Synthèse de l'analyse bivarié

L'analyse bivariée suggère que la dépression est surtout associée à des facteurs de contexte et de mode de vie : les étudiants dépressifs présentent un **stress financier plus élevé**, une **charge travail/études plus importante**, une proportion plus forte de dépression chez les profils **alimentaires "Unhealthy"**, et chez ceux qui dorment **moins de 5 heures**.

À l'inverse, la **satisfaction des études** varie peu selon la dépression et selon le **diplôme**, ce qui laisse penser qu'elle discrimine faiblement les groupes dans cette base. Ces tendances devront être **confirmées** par les tests d'inférence (comparaison et indépendance).

III. REPONSES AUX QUESTIONS DU PROJET

1. DETERMINONS L'INTERVAL DE CONFIANCE LA PROPORTION D'ETUDIANTS AYANT DEJA EU DES PENSEES SUICIDAIRES

```
[1] "Intervalle de confiance pour la proportion d'étudiants ayant déjà eu des pensées suicidaires:"

   No   Yes
10245 17656
[1] "Proportion estimée à :"
[1] 0.6328089
[1] "Intervalle de confiance (95 %):"
[1] 0.6271212 0.6384689
attr(,"conf.level")
[1] 0.95
```

On estime qu'environ **63 %** des étudiants interrogés ont déjà eu des pensées suicidaires. L'intervalle de confiance à **95 %** est compris entre : **62.7 %** et **63.8 %**.

On retient que ce résultat met en lumière une prévalence inquiétante des pensées suicidaires chez les étudiants. Il s'agit d'un facteur déterminant à prendre en compte dans l'analyse de la santé mentale étudiante.

2. MOYENNE ET MEDIANE

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

2.1.1 Estimons la moyenne des heures de travail ou d'études pour les étudiants souffrant de dépression.

```
Moyenne des heures de travail/étude (étudiants dépressifs) : 7.81 h
[1] " Nous obtenons comme interval de confiance avec un niveau de confiance de 95%.:"
[1] 7.754689 7.860516
attr(,"conf.level")
[1] 0.95
```

On estime que les étudiants **dépressifs** consacrent en moyenne **7,81 heures par jour** au travail/études. L'**intervalle de confiance à 95 %** pour cette moyenne est compris entre **7,75 h** et **7,87 h**.

Ce résultat suggère que, dans cette population, la charge quotidienne des étudiants dépressifs est **élevée** et surtout estimée avec une **grande précision** (intervalle très resserré, probablement dû à un effectif important).

2.1.1 Estimons la médiane des heures de travail ou d'études pour les étudiants souffrant de dépression.

```
Médiane des heures travail/étude (dépressifs) : 9.00 h
IC 95% (bootstrap) : [9.00 ; 9.00]
```

On estime que la **médiane** des heures de travail/étude chez les étudiants **dépressifs** est de **9,00 h**. L'**intervalle de confiance à 95 %** obtenu par **bootstrap** est **[9,00 h ; 9,00 h]**.

Ce résultat indique que, dans cette population, **la valeur centrale** du temps de travail/étude des étudiants dépressifs est **élevée** et surtout estimée avec une **très grande précision**. Un intervalle aussi resserré s'explique généralement par un **effectif important** et par une variable **discrète** (heures entières) dont la médiane est très stable d'un échantillon bootstrap à l'autre.

2.2 Estimons la moyenne et la médiane du stress financier pour les étudiants avec et sans dépression.

2.2.1 Estimons la moyenne du stress financier pour les étudiants avec et sans dépression.

```
la moyenne du stress financier pour les étudiants avec et sans dépression: 2.52 h
[1] " Nous obtenons comme interval de confiance avec un niveau de confiance de 95%.:"
[1] 2.494257 2.543356
attr(,"conf.level")
[1] 0.95
```

On estime que la **moyenne du stress financier** (dans l'échantillon analysé) est de **2,52**. L'**intervalle de confiance à 95 %** pour cette moyenne est compris entre **2,49** et **2,54**.

Ce résultat indique un niveau moyen de stress financier **plutôt modéré**, et surtout estimé avec une **très grande précision** (intervalle extrêmement resserré, cohérent avec un effectif important). Attention toutefois : tel qu'affiché, ce chiffre semble être une **moyenne globale** ; si ton objectif est bien de comparer **avec vs sans dépression**, il faut produire **deux moyennes et deux IC** (un par groupe), sinon on ne peut pas conclure sur une différence entre groupes

2.2.2 Estimons la médiane du stress financier pour les étudiants avec et sans dépression.

```
Médiane des heures de travail/étude (étudiants dépressifs) : 2.00 h

        Bootstrap

data:  bd$stress_financier
1000 replicates

95 percent confidence interval:
 2 2
sample estimates:
original value
             2
```

On estime que la **médiane du stress financier** (dans l'échantillon analysé) est de **2,00**. L'**intervalle de confiance à 95 %**, obtenu par **bootstrap** (1000 réplications), est **[2 ; 2]**.

Ce résultat indique que le niveau "typique" de stress financier est **plutôt faible à modéré** et surtout estimé avec une **très grande précision**. Un intervalle aussi resserré est cohérent avec une variable **discrète** (échelle 1–5) et une médiane très stable à travers les rééchantillonnages.

3. DIFFERENCE DE MOYENNE

3.1 Vérifions si la satisfaction des études diffère significativement entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas

D'abord vérifions la normalité par le test de shapiro la distribution de satisfaction_etudes chez les étudiants dépressifs et la distribution de satisfaction_etudes chez les non dépressifs

H0 : la distribution est normale
H1 : La distribution n'est pas normale

```
        Shapiro-Wilk normality test

data:  sample(x_oui, min(length(x_oui), 5000))
W = 0.89595, p-value < 2.2e-16

Interprétation : On rejette l'hypothèse de normalité pour la satisfaction_etude(oui_dépressif) (p =
7.005655e-50 )


        Shapiro-Wilk normality test

data:  sample(x_non, min(length(x_non), 5000))
W = 0.8975, p-value < 2.2e-16

Interprétation : On rejette l'hypothèse de normalité pour la satisfaction_etude(non_dépressif) (p =
1.259773e-49 )
```

Groupe "oui" (dépressifs) : on obtient **W = 0,89595** avec une **p-value extrêmement petite** (affichée < 2,2e-16, et le script la donne ≈ 7,0e-50). On **rejette donc l'hypothèse de normalité** : la distribution de satisfaction_etudes chez les étudiants dépressifs n'est pas normale.

Groupe "non" (non dépressifs) : même constat, **W = 0,8975** et **p-value extrêmement petite** (≈ 1,26e-49). On **rejette aussi l'hypothèse de normalité** : la distribution de satisfaction_etudes chez les non dépressifs n'est pas normale.

On tire que : comme satisfaction_etudes est une échelle **discrète/ordinale (1–5)** et que la normalité est rejetée dans les deux groupes, on a un argument propre pour **éviter un t-test** et privilégier un test **robuste/non paramétrique** (ex. **Wilcoxon/Mann-Whitney**) pour comparer les deux groupes.

Test non paramétrique de wilcoxon

H0 : le niveau de satisfaction des études ne diffère pas entre étudiants dépressifs et non dépressifs (distributions identiques).

H1 : le niveau de satisfaction des études diffère entre étudiants dépressifs et non dépressifs.

```
        Wilcoxon rank sum test with continuity correction

data:  depression_oui and depression_non
W = 76236987, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

Interprétation : On rejette H0. Il existe une différence significative de satisfaction_etudes entre étudiants dépressifs et non (p = 1.362006e-173 )
```

Le test de **Wilcoxon rank-sum** (Mann–Whitney) compare la distribution de satisfaction_etudes entre les étudiants **dépressifs** et **non dépressifs**.

Ici, la **p-value < 2,2e-16** (et le script affiche p ≈ $\mathbf{1{,}36\times10^{-173}}$) : c'est écrasant. Donc on **rejette H0**. Autrement dit, le niveau de satisfaction des études **diffère significativement** entre les deux groupes.

3.2 Vérifions si les niveaux de satisfaction au travail diffèrent-ils significativement selon le diplôme suivi

Ici la variable diplôme_suivi présente plusieurs modalité donc le test le mieux indiqué est l'**ANOVA** mais nous ne pouvons pas l'appliquer ici raisons :

la variable satisfaction_au_travail est une variable ordinale 1–5
beaucoup de groupes
grand n ⇒ tests de normalité rejettent quasi toujours

Donc tu justifies proprement : **Kruskal–Wallis** (et post-hoc Dunn) est plus cohérent.

Test non paramétrique de Kruskal-Wallis

H0 : la distribution (et donc la position centrale/médiane) de satisfaction_etudes est la même pour tous les niveaux de diplome_suivi.

H1 : au moins un diplôme a une distribution (ou une médiane) de satisfaction_etudes différente.

```
        Kruskal-Wallis rank sum test

data:  satisfaction_etudes by diplome_suivi
Kruskal-Wallis chi-squared = 156.71, df = 27, p-value < 2.2e-16

Interprétation : On rejette H0. La satisfaction_etudes diffère significativement selon le diplôme suivi
(p = 3.073921e-20 )
```

Le test de **Kruskal–Wallis** indique une différence nette de satisfaction_etudes selon diplome_suivi.

On obtient $\boldsymbol{\chi^2 = 156{,}71}$ avec **27 degrés de liberté** et une **p-value** $\mathbf{< 2{,}2\times10^{-16}}$ (le script l'affiche aussi ≈ $\mathbf{3{,}07\times10^{-20}}$). Donc on **rejette H0** : la distribution (et donc le niveau central) de satisfaction des études **n'est pas la même** pour tous les diplômes.

Ce résultat dit une chose précise : **au moins un diplôme** se distingue des autres en termes de satisfaction.

4. INDEPENDANCE

4.1 Vérifions si La dépression est indépendante des habitudes alimentaires (saines/modérées)

Nous avons deux variables qualitatives , le test approprié ici est celui de Chi-deux d'indépendance

Test du Chi-deux d'indépendance

H0 (hypothèse nulle) : la dépression est indépendante des habitudes alimentaires
Autrement dit, la proportion d'étudiants dépressifs est la même quel que soit le type d'habitudes alimentaires (Healthy / Moderate / Unhealthy).

H1 (hypothèse alternative) : la dépression n'est pas indépendante des habitudes alimentaires
Autrement dit, il existe une association : la proportion d'étudiants dépressifs varie selon les habitudes alimentaires.

```
        Pearson's Chi-squared test

data:  contingency_table
X-squared = 1202.3, df = 2, p-value < 2.2e-16

Interprétation : On rejette H0. Il existe une association significative entre habitudes_alimentaires et
depression (p = 8.465054e-262 )
```

Le **test du Chi-deux de Pearson** appliqué au tableau habitudes_alimentaires × depression donne :

- $\chi^2 = 1202.3$, **ddl = 2**
- **p-value < 2.2e-16** (dans ton affichage : $p = 8.47 \times 10^{-262}$, donc pratiquement 0)

Conclusion nette : **on rejette H0**.

Il existe une **association statistiquement significative** entre les habitudes alimentaires et l'état dépressif : **la proportion d'étudiants dépressifs n'est pas la même** selon que l'alimentation est *Healthy*, *Moderate* ou *Unhealthy*.
Point important : avec **27 901** observations, une p-value aussi minuscule dit "il y a un lien", mais **ne dit pas si le lien est fort**. Pour juger l'intensité réelle, regardons **Cramer's V** (avec assocstats(contingency_table))

Calcul du coefficient de Cramer (V)

```
                   X^2 df P(> X^2)
Likelihood Ratio 1218.2  2        0
Pearson          1202.3  2        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.203
Cramer's V        : 0.208
Interprétation (Cramer's V) : V = 0.208
Force de l'association : faible à modérée.
```

Le test du Chi-deux met en évidence une association statistiquement significative entre habitudes alimentaires et dépression ($p \approx 0$). Toutefois, l'intensité de cette association reste **faible à modérée** (Cramer's V = 0.208), ce qui suggère que les habitudes alimentaires contribuent à différencier les états dépressifs, sans constituer à elles seules un déterminant dominant.

4.2 Vérifions si la durée du sommeil (par exemple, moins de 5 heures, 5-6 heures, 7-8 heures) est-elle indépendante de la dépression.

Nous avons deux variables qualitatives , le test approprié ici est celui de Chi-deux d'indépendance

Test du Chi-deux d'indépendance

H0 (hypothèse nulle) :
La dépression est indépendante de la durée du sommeil.
Autrement dit, la proportion d'étudiants dépressifs (oui/non) est la même quelle que soit la catégorie de durée de sommeil (5–6h, 7–8h, <5h, >8h).

H1 (hypothèse alternative) :
La dépression n'est pas indépendante de la durée du sommeil.
Autrement dit, il existe une association : la proportion d'étudiants dépressifs varie selon la durée du sommeil.

```
        Pearson's Chi-squared test

data:  contingency_tab
X-squared = 276.32, df = 3, p-value < 2.2e-16


Interprétation : On rejette H0. Il existe une association statistiquement significative entre
duree_sommeil et depression (p = 1.326841e-59 ).
Conditions : effectifs attendus >= 5 -> Chi-2 pertinent.
```

On observe une association statistiquement significative entre la durée du sommeil et l'état

dépressif des étudiants ($\chi^2 = 276{,}32$; ddl = 3 ; $p < 2{,}2\times10^{-16}$). L'hypothèse d'indépendance est donc clairement rejetée.

Cela signifie que la proportion d'étudiants dépressifs varie selon les catégories de durée du sommeil (5–6 h, 7–8 h, < 5 h, > 8 h). Autrement dit, la durée du sommeil constitue un facteur associé à la dépression chez les étudiants. Les conditions d'application du test étant respectées (effectifs attendus ≥ 5), ce résultat est statistiquement fiable et pertinent. Pour juger l'intensité réelle de cette association, regardons **Cramer's V** (avec assocstats(contingency_tab)

Calcul du coefficient de Cramer (V)

```
                  X^2 df P(> X^2)
Likelihood Ratio 276.37  3        0
Pearson          276.32  3        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.099
Cramer's V        : 0.1
Interprétation (Cramer's V) : V = 0.1
Force de l'association : très faible (même si la p-value peut être significative avec un grand
effectif).
```

Le test du Chi-deux (Pearson et Likelihood Ratio) confirme l'existence d'une association statistiquement significative entre la durée du sommeil et l'état dépressif ($\chi^2 \approx 276$; ddl = 3 ; $p \approx 0$). Autrement dit, on rejette formellement l'hypothèse d'indépendance.
Cependant, la **taille de l'effet est très faible**. Le coefficient de Cramér (V = 0,10) indique que, même si l'association est statistiquement détectable, son **intensité est limitée sur le plan pratique**. Cette significativité est en grande partie liée à la taille très élevée de l'échantillon, qui rend le test extrêmement sensible à de faibles écarts.

En résumé :

- **Statistiquement**, la durée du sommeil et la dépression sont liées.
- **Substantiellement**, cette relation est faible et ne suffit pas, à elle seule, à expliquer l'état dépressif.

La durée du sommeil agit comme un **facteur associé**, mais non comme un déterminant majeur isolé.

CONCLUSION GENERALE

Rappel de la problématique

Ce travail avait pour objectif d'analyser les déterminants associés à l'état dépressif chez les étudiants, en mettant en relation la dépression avec plusieurs dimensions clés de leur vie académique et quotidienne : charge de travail et d'étude, satisfaction des études, stress financier, habitudes alimentaires et durée du sommeil. L'enjeu était de comprendre quels facteurs sont statistiquement liés à la dépression, et dans quelle mesure ces relations sont réellement pertinentes au-delà de la seule significativité statistique.

Résumé des principaux résultats obtenus

Les analyses descriptives et inférentielles ont mis en évidence plusieurs constats majeurs.
Premièrement, les étudiants dépressifs présentent une charge de travail/études élevée, estimée avec une grande précision, suggérant une pression académique importante dans cette population. La satisfaction des études diffère significativement entre étudiants dépressifs et non dépressifs, confirmant un lien fort entre mal-être psychologique et perception négative du parcours académique.
Par ailleurs, le stress financier apparaît également associé à la dépression, bien que la médiane observée soit relativement stable, traduisant une vulnérabilité partagée mais non exclusive aux étudiants dépressifs.
Les tests d'indépendance ont montré des associations statistiquement significatives entre la dépression et les habitudes alimentaires ainsi qu'entre la dépression et la durée du sommeil. Toutefois, les mesures de taille d'effet (Cramér's V) indiquent que ces associations sont faibles à très faibles, suggérant que ces facteurs jouent un rôle secondaire pris isolément.

Limites de l'analyse
Plusieurs limites doivent être soulignées.
Tout d'abord, la très grande taille de l'échantillon rend les tests statistiques extrêmement sensibles, ce qui conduit à détecter des associations significatives même lorsque leur impact réel est faible. Ensuite, l'analyse repose sur des données déclaratives, exposées à des biais de perception et de déclaration. De plus, la nature transversale des données empêche toute interprétation causale : il n'est pas possible de déterminer si les facteurs étudiés sont des causes ou des conséquences de la dépression. Enfin, certaines variables catégorielles larges (par exemple les durées de sommeil ou les habitudes alimentaires) peuvent masquer des nuances importantes au sein des groupes.

Perspectives et pistes d'approfondissement
Pour améliorer la portée de ces résultats, plusieurs prolongements sont envisageables. Une modélisation multivariée (régression logistique) permettrait d'évaluer l'effet propre de chaque facteur en contrôlant les autres variables. L'intégration d'indicateurs psychosociaux supplémentaires (soutien social, charge émotionnelle, conditions de vie) renforcerait la compréhension globale du phénomène. Une analyse longitudinale, si les données le permettent, offrirait une lecture dynamique de l'évolution de la dépression dans le temps. Enfin, une segmentation des profils d'étudiants pourrait aider à identifier des groupes à risque spécifiques et à orienter plus efficacement les actions de prévention.

ANNEXE

Codes d'analyse en R utilisés pour les traitements des données

```
# *Importation de la base de données et inspection*
```{r}
Sd <- read.csv("C:/Users/HP/Desktop/INSSEDS/PROJET STAT INF/PROJET STAT INF R/Student_Depression.csv",
stringsAsFactors=TRUE)
View(Sd)
print(Sd)
```
```

```
*Encodage de variable*

```{r}
# Transformation de variables en factor
Sd$id <- factor(Sd$id)

# Transformation de variables en factor avec encodage
Sd$depression <- factor(Sd$depression, levels = c(0, 1), labels = c("non", "oui"))

# Si vous voulez garder pression liee au travail comme factor sans encodage sémantique
Sd$pression_liee_au_travail <- factor(Sd$pression_liee_au_travail)

# Si vous voulez garder satisfaction_travail comme factor sans encodage sémantique
Sd$satisfaction_travail <- factor(Sd$satisfaction_travail)

head(Sd)
```

```
*Structure de la base données*
```{r}
str(Sd)
```
```

```
#2- Identification des doublons

```{r}
doublon=duplicated(Sd$id)
sum(doublon)
```
```

```
3-1. Vérification des valeurs manquantes

```{r}
library(VIM)
# Graphique avec pourcentages en visuel
aggr(Sd,
     col = c("navyblue", "red"),
     numbers = TRUE,
     sortVars = TRUE,
     cex.axis = 0.7,
     gap = 3,
     ylab = c("Proportion de valeurs manquantes", "les proportions"),
     # Options pour mieux afficher les pourcentages
     cex.numbers = 0.8,        # Taille des nombres
     prop = TRUE               # Affiche en proportions (pourcentages)

)

# Nombre de données manquantes
sum(!complete.cases(Sd))
```
```

```
3-2. Traitement des valeurs manquantes

Imputation par la médiane

```{r}
#Calcule de la médiane
median(Sd$stress_financier,na.rm = TRUE)

#Imputation par la médiane
Sd$stress_financier[is.na(Sd$stress_financier)]=median(Sd$stress_financier,na.rm = TRUE)
aggr(Sd, col = c("navyblue", "red"), numbers = TRUE, sortVars = TRUE, cex.axis = 0.7,
     gap = 3, ylab = c("Proportion de valeurs manquantes", "les proportions")) #graphique
sum(!complete.cases(Sd))  # nombre de données manquantes
```
```
```

```
*4-2. Affichage de la boite à moustache*
```{r}
# Charger les bibliothèques nécessaires
library(VIM)
library(ggplot2)

afficher_boites_a_moustache <- function(dataframe) {
  # Sélectionner uniquement les colonnes numériques
  colonnes_numeriques <- dataframe[, sapply(dataframe, is.numeric), drop = FALSE]

  if (ncol(colonnes_numeriques) > 0) {
    # Reshape des données manuellement pour ggplot2
    dataframe_melted <- stack(colonnes_numeriques)
    colnames(dataframe_melted) <- c("Valeur", "Variable")

    # Créer un graphique de boites à moustaches avec ggplot2
    ggplot(dataframe_melted, aes(x = Variable, y = Valeur, fill = Variable)) +
      geom_boxplot(outlier.color = "red", outlier.shape = 16, outlier.size = 2, alpha = 0.7) +
      labs(title = "Boites à Moustaches des Variables Numériques", x = "Variables", y = "Valeurs
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
            axis.text.y = element_text(size = 10),
            plot.title = element_text(hjust = 0.5, size = 14, face = "bold")) +
      scale_fill_manual(values = rainbow(ncol(colonnes_numeriques))) +
      geom_vline(xintercept = 1:ncol(colonnes_numeriques), color = "gray", linetype = "dotted")
```

```
# 4- Traitement des valeurs extremes et/ou aberrantes
```{r}
# Charger les bibliothèques nécessaires
library(VIM)
library(ggplot2)

# Définir la fonction pour afficher les boites à moustaches après winsorisation
afficher_boites_a_moustache_winsorisee <- function(dataframe, lower_quantile = 0.05, upper_quantile
0.95) {
  # Sélectionner uniquement les colonnes numériques
  colonnes_numeriques <- dataframe[, sapply(dataframe, is.numeric), drop = FALSE]

  if (ncol(colonnes_numeriques) > 0) {
    # Appliquer la winsorisation manuelle à chaque colonne numérique
    colonnes_numeriques_winsor <- as.data.frame(
      lapply(colonnes_numeriques, function(col) {
        # Calculer les quantiles inférieur et supérieur
        q_lower <- quantile(col, lower_quantile, na.rm = TRUE)
        q_upper <- quantile(col, upper_quantile, na.rm = TRUE)

        # Appliquer la winsorisation
        col[col < q_lower] <- q_lower
        col[col > q_upper] <- q_upper
        return(col)
      })
```

```
1 - *VARIABLE QUANTITATIVES*

# Resumé statistique des variables quantitatives

```{r}
# Chargement des packages nécessaires
library(dplyr)
library(moments)  # pour skewness et kurtosis

resume_numerique <- function(df) {
  # Sélection des variables quantitatives (numériques)
  num_vars <- df %>%
    select(where(is.numeric))

  if (ncol(num_vars) == 0) {
    cat("Aucune variable quantitative trouvée.\n")
    return(invisible(NULL))
  }
```

```
# Histogramme des variables quantitatifs

```{r}
afficher_histogrammes <- function(df, exclude_cols = c('id'), color = 'steelblue', n_cols = 3) {

  # Charger dplyr pour utiliser %>%
  library(dplyr)

  # Extraction des variables quantitatives
  quant_vars <- df %>%
    select(where(is.numeric)) %>%
    names()

  quant_vars <- quant_vars[!quant_vars %in% exclude_cols]

  if (length(quant_vars) == 0) {
    cat("⚠️ Aucune variable quantitative trouvée.\n")
    return(invisible(NULL))
  }

  # Ajustement dynamique du nombre de colonnes
  n_vars <- length(quant_vars)
```

```
# Diagramme en Barre des variables qualitatives

```{r}
library(ggplot2)

angoran_ql_graph <- function(facteur) {
  # Création d'un data frame contenant les fréquences absolues et relatives de chaque modalité
  df <- data.frame(table(facteur))
  colnames(df) <- c("Modalite", "Freq")
  df$freq_relatives <- round(100 * df$Freq / sum(df$Freq), 2)

  # Diagramme en barre horizontal avec les fréquences relatives
  ggplot(df, aes(x = reorder(Modalite, freq_relatives), y = freq_relatives, fill = Modalite)) +
    geom_bar(stat = "identity", alpha = 0.8, width = 0.7) +
    geom_text(aes(label = paste0(freq_relatives, "%")),
              position = position_stack(vjust = 0.5),
              color = "black",
              fontface = "bold") +
    labs(

      x = "Modalités",
      y = "Fréquences relatives (%)"
    ) +
    theme_minimal() +
    theme(
```

```
# Heure de travaille en fonction des  de l'etat  dépressifs des etudiants

```{r}
ggplot(Sd, aes(x = depression, y = nombre_heure_travail_etude, fill = depression)) +
  geom_boxplot() +
  labs(title = "Nombre d'heure d'étude en fonction de la dépression ", x = "Depression", y = "Nombre heur
travail etude") +
  theme_minimal()
```
```

```
Depression - Habitude alimentaire

```{r}
# Chargement du package nécessaire
library(ggplot2)
library(scales)  # pour formater les pourcentages

# Graphique de répartition en % par dépression et habitudes alimentaires
repartition <- ggplot(Sd, aes(x = habitudes_alimentaires, fill = depression)) +
  geom_bar(position = "fill", color = "black") +
  scale_y_continuous(labels = percent_format()) +  # pour affichage en pourcentage
  geom_text(stat = "count",
            aes(label = scales::percent(..count../tapply(..count.., ..x.., sum)[..x..], accuracy = 1)),
            position = position_fill(vjust = 0.5),
            size = 3, color = "white") +
  labs(
    title = "Répartition par état dépressif et par tranche d'âge ",
    x = "Habitude alimentaire",
    y = "Proportion (%)",
    fill = "Etat depressif"
  ) +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal()

# Affichage
print(repartition)
```
```

```
ESTIMATION ET TESTS STATISTIQUES

1 Intervalle de confiance pour la proportion d'étudiants ayant déjà eu des pensées suicidaires


```{r}
# Effectif pensée suicidaire ou non
table(Sd$pensees_suicidaire)
```
```
```

```r
# Effectif pensee suicidaire ou non
print("Intervalle de confiance pour la proportion d'étudiants ayant déjà eu des pensées suicidaires:")
table(Sd$pensees_suicidaire)

# L'intervalle de confiance pour la proportion d'étudiants ayant déjà eu des pensées suicidaire 17656 sur
27901 etudiants est:
# Condition : echec et succes > 10 ( vérifiée)
nb_succes <- 17656
nb_essais <- 27901
nb_echecs <- nb_essais - nb_succes

print("Proportion estimée à :")

nb_succes / nb_essais

print("Intervalle de confiance (95 %):")

# Interval de confiance
binom.test(nb_succes,nb_essais )$conf.int
```

```r
# Calcul de la moyenne des heures de travail/étude pour les étudiants dépressifs
moyenne_heures <- mean(dframe$nombre_heure_travail_etude)

# Affichage avec deux décimales
cat(sprintf("Moyenne des heures de travail/étude (étudiants dépressifs) : %.2f h\n", moyenne_heures))

# Puisque la taille de l'échantillon est superieur à 30 elle est suffisament grande pour qu'on la calcule
directement :

print(" Nous obtenons comme interval de confiance avec un niveau de confiance de 95%.:")
t.test(dframe$nombre_heure_travail_etude ,conf.int=TRUE,conf.level = 0.95)$conf.int
```

```r
*a 1 -Estimation de la mediane des heures de travail ou d'études pour les étudiants souffrant de
depression.*

```{r}
library(RVAideMemoire)

x <- Sd$nombre_heure_travail_etude[Sd$depression == "oui"]
x <- x[!is.na(x)]

med <- median(x)

set.seed(1)
boot_med <- bootstrap(x, function(x, i) median(x[i]), nrep = 5000)

cat(sprintf("Médiane des heures travail/étude (dépressifs) : %.2f h\n", med))
cat(sprintf("IC 95%% (bootstrap) : [%.2f ; %.2f]\n",
            boot_med$conf.int[1], boot_med$conf.int[2]))
```

```r
# Test de Wilcoxon (Mann-Whitney) : satisfaction_etudes selon l'état dépressif
# H0 : Les niveaux de satisfaction (distributions / médianes) sont identiques entre les deux groupes
# H1 : Les niveaux de satisfaction sont différents entre les deux groupes

depression_oui <- df_sd$satisfaction_etudes[df_sd$depression == "oui"]
depression_non <- df_sd$satisfaction_etudes[df_sd$depression == "non"]

# Suppression des valeurs manquantes
depression_oui <- depression_oui[!is.na(depression_oui)]
depression_non <- depression_non[!is.na(depression_non)]

wilcoxon_test <- wilcox.test(depression_oui, depression_non, alternative = "two.sided")
print(wilcoxon_test)

# Interprétation
if(wilcoxon_test$p.value > 0.05) {
  cat("Interprétation : On ne rejette pas H0. Il n'y a pas de différence significative de
satisfaction_etudes entre étudiants dépressifs et non (p =",
      format(wilcoxon_test$p.value, scientific = TRUE), ")\n")
} else {
  cat("Interprétation : On rejette H0. Il existe une différence significative de satisfaction_etudes entr
étudiants dépressifs et non (p =",
      format(wilcoxon_test$p.value, scientific = TRUE), ")\n")
}
```

```r
# Kruskal-Wallis : satisfaction_etudes ~ diplome_suivi
df_sds <- Sd[, c("satisfaction_etudes", "diplome_suivi")]
df_sds <- na.omit(df_sds)

kw_test <- kruskal.test(satisfaction_etudes ~ diplome_suivi, data = df_sds)
print(kw_test)

# Interprétation
if(kw_test$p.value > 0.05) {
  cat("Interprétation : On ne rejette pas H0. Aucun écart significatif de satisfaction_etudes selon le
diplôme suivi (p =",
      format(kw_test$p.value, scientific = TRUE), ")\n")
} else {
  cat("Interprétation : On rejette H0. La satisfaction_etudes diffère significativement selon le diplôme
suivi (p =",
      format(kw_test$p.value, scientific = TRUE), ")\n")
}
```
```

```r
# Test du Chi-deux
chi_squared_test <- chisq.test(contingency_table)
print(chi_squared_test)

# Interprétation
if(chi_squared_test$p.value > 0.05) {
  cat("Interprétation : On ne rejette pas H0. Aucune association significative entre
habitudes_alimentaires et depression (p =",
      format(chi_squared_test$p.value, scientific = TRUE), ")\n")
} else {
  cat("Interprétation : On rejette H0. Il existe une association significative entre
habitudes_alimentaires et depression (p =",
      format(chi_squared_test$p.value, scientific = TRUE), ")\n")
}

# Vérification des conditions d'application du Chi-deux (effectifs attendus)
if(any(chi_squared_test$expected < 5)) {
  cat("Avertissement : certains effectifs attendus sont < 5. Le test du Chi-deux peut être moins fiable
;\n",
      "envisagez fisher.test(contingency_table) ou regroupez des modalités.\n")
}
```

```r
# Charger le package (installer une seule fois si besoin)
# install.packages("vcd")
library(vcd)

# Statistiques d'association
assoc_results <- assocstats(contingency_tab)
print(assoc_results)

# Interprétation rapide de Cramer's V (force du lien)
V <- as.numeric(assoc_results$cramer)

cat("Interprétation (Cramer's V) : V =", round(V, 3), "\n")

if(V < 0.10) {
  cat("Force de l'association : très faible (même si la p-value peut être significative avec un grand
effectif).\n")
} else if(V < 0.30) {
  cat("Force de l'association : faible à modérée.\n")
} else if(V < 0.50) {
  cat("Force de l'association : modérée à forte.\n")
} else {
  cat("Force de l'association : forte.\n")
}
```