MINISTAIRE DE L'ENNSEIGNEMENT SUPRIEUR

ET DE LA RECHERCHE SCIENTIFIQUE

REPUBLIQUE COTE D'IVOIRE

UNION-DISCIPLINE-TRAVAIL

MASTER 1

INGENIEUR STATISTICIEN -DATA ANALYSTE-
DATA SCIENCE

MINI PROJET :

ANALYSE DE LA SERIE TEMPORELLE DES VENTES DE LA CHAINE D'EPICERIE Corporación Favorita

ETUDIANT

KOUAHON ESTELLE

PROFESSEUR

AKPOSSO MARTIAL

SOMMAIRE

INTRODUCTION

PARTIE 1 **6**

DICTIONNAIRE DE DONNEES **7**

IMPORTATION DU JEU DE DONNEES 7

Traitement des valeurs manquantes 8

Traitement des valeurs extrêmes et aberrantes 8

VISUALISATION AVANT TRAITEMENT 9

VISUALISATION APRES TRAITEMENT 10

TRAITEMENT DES DOUBLONS 10

CATEGORISATION DES MODALITES DES VARIABLES 10

PARTIE2 : ANALYSE UNIVARIEE 11

INTERPRETATION 11

Diagramme en bande diplôme suivi 11

Diagramme en camembert 12

STATISTIQUE BIVARIEE 13

Tableau de contingence ville et dépression 13

Croisement pensees_suicidaire et dépression 14

Croisement diplôme suivi et dépression 15

Croisement habitude_alimentaire et dépression 15

Croisement sexe et depression 16

Croisement antecedant_familiaux_maladie_mentale et dépression 17

Croisement duree_sommeil et dépression 18

Croisement age et dépression 19

PARTIE 3 21

INTERVALLE DE CONFIANCE 21

IDENTIFICATION TEST STATISTIQUE (Question 2) 22

Test d'ajustement à une loi 22

Visualisation de moyenne et de médiane des populations 22

Visualisation de test ajustement stress_financier et depression _____23
Test de significativité_____23
Visualisation de moyenne et de médiane des populations_____24
Visualisation test de conformité nombre_heure_travail_etude et depression_____25
Test de significativité_____25
tableau de contingences_____26
estimation des medianes_____26
TEST D'INDEPENDANCE (QUESTION 4)_____27
visualisations tests d'indépendances_____27
TEST DE COMPARAISON DE MEDIANES (Question 3)_____28
CONCLUSION GENERALE_____29
RECOMANDATIONS_____29
CODES SOURCE_____30

AVANT-PROPOS

La santé mentale des étudiants constitue un enjeu majeur dans le paysage éducatif contemporain. Face à des pressions académiques, financières et sociales croissantes, de nombreux étudiants sont confrontés à des défis psychologiques, dont la dépression, qui peut avoir des répercussions profondes sur leur bien-être, leur performance académique et leur qualité de vie. Cette étude vise à explorer les facteurs influençant la dépression chez les étudiants en utilisant des méthodes statistiques robustes, afin de mieux comprendre les dynamiques sous-jacentes et de proposer des pistes d'intervention préventive.

L'analyse des données relatives à la santé mentale des étudiants permet d'identifier les tendances et les facteurs explicatifs associés à la dépression. Ce projet s'appuie sur un ensemble de données détaillé, comprenant des informations démographiques, des indicateurs de performance académique, des habitudes de vie, des antécédents de santé mentale et des réponses à des échelles de dépression standardisées. Ces données offrent une opportunité unique d'étudier les corrélations entre divers facteurs et la prévalence de la dépression, tout en fournissant des insights précieux pour les chercheurs en psychologie, en science des données et en éducation.

L'objectif principal de cette étude est d'analyser l'association entre les facteurs liés au mode de vie, aux études et la dépression chez les étudiants. Les résultats permettront de formuler des recommandations fondées sur des données probantes pour améliorer la santé mentale des étudiants et concevoir des stratégies d'intervention précoce. Les tâches spécifiques incluent l'estimation de la prévalence des pensées suicidaires, la comparaison des indicateurs de performance académique et de stress financier entre étudiants dépressifs et non dépressifs, ainsi que l'évaluation de l'indépendance entre la dépression et des facteurs tels que les habitudes alimentaires ou la durée du sommeil.

Ce rapport présente une analyse approfondie des données, en mettant l'accent sur les méthodes statistiques utilisées et les résultats obtenus. Les livrables incluent une quantification des facteurs les plus influents sur la dépression, des recommandations pour améliorer la santé mentale des étudiants, et des visualisations clés pour faciliter l'interprétation des résultats. Nous espérons que cette étude contribuera à une meilleure compréhension des enjeux de santé mentale chez les étudiants et à la mise en place de politiques éducatives et de soutien adaptées.

Introduction

La santé mentale des étudiants, en particulier la dépression, est un enjeu critique dans le monde académique. Les pressions académiques, financières et sociales peuvent avoir un impact profond sur leur bien-être, affectant leur performance et leur qualité de vie.

Problématique

Malgré une prise de conscience croissante, les facteurs spécifiques influençant la dépression chez les étudiants restent mal compris. Quels sont les éléments liés au mode de vie, aux études ou aux antécédents personnels qui jouent un rôle significatif dans l'apparition de la dépression ?

Objectif Général

Ce projet vise à étudier les facteurs influençant la dépression chez les étudiants en utilisant des méthodes statistiques, afin d'identifier les corrélations et tendances significatives.

Objectifs Spécifiques

1. Analyser la prévalence des pensées suicidaires.
2. Comparer les indicateurs de performance académique et de stress financier entre étudiants dépressifs et non dépressifs.
3. Évaluer les différences de satisfaction liées aux études et au travail.
4. Vérifier l'indépendance entre la dépression et des facteurs tels que les habitudes alimentaires ou la durée du sommeil.

Résultats Attendus

- Une quantification des facteurs les plus influents sur la dépression.
- Des recommandations pour améliorer la santé mentale des étudiants.

Méthodologie

Python est le logiciel utilisé pour l'analyse statistique avancée et la visualisation des données.Cette approche combinée permettra une analyse robuste et une communication claire des résultats, contribuant à une meilleure compréhension des déterminants de la dépression chez les étudiants.

PARTIE 1 : ANALYSE EXPLOORATOIRE DES DONNEES

Le prétraitement des données est une étape essentielle dans tout projet d'analyse statistique. Il permet de s'assurer que les données sont propres, cohérentes et prêtes à etre analysées. Il couvre trois étapes : l'identification et la gestion des valeurs manquantes, détection des valeurs aberrantes et extrêmes et la recherche et la suppression des données.

1- DICTIONNAIRE DE DONNEES

Avant de procéder à l'analyse des données, il est crucial de comprendre la structure et la signification de chaque variable. Ce dictionnaire des données fournit une description détaillée de chaque variable, en précisant son type (numérique, catégoriel, binaire), sa nature (qualitative ou quantitative), et sa signification. Cette description permet de guider les analyses statistiques et de garantir une interprétation correcte des résultats.

CARACTERISTIQUES	VARIABLES	TYPE	DESCRIPTION
SOCIODEMOGRAPHIQUES	**Id**	**numerique**	**Identifiant unique de chaque étudiant**
	Sexe	**Catégoriel**	**Sexe de l'étudiant**
	Age	**Numérique**	**Âge de l'étudiant en années**
	Ville	**Catégoriel**	**Ville de résidence de l'étudiant**
	diplome_suivi	Catégoriel	Diplôme en cours ou obtenu
	profession	**Catégoriel**	**Profession principale de l'étudiant (étudiant à temps plein, salarié, etc.)**
ÉTUDES ET TRAVAIL	**pression_academique**	**Ordinale**	**Niveau de pression ressenti lié aux études**
	pression_liee_au_travail	**Ordinal**	**Niveau de pression ressenti lié au**

			travail
	moyenne_notes	Ordinal	Moyenne générale des notes de l'étudiant
	nombre_heure_travail_etude	Numerique	Nombre total d'heures consacrées au travail et aux études par jour
	satisfaction_travail	Ordinal	Niveau de satisfaction par rapport au travail
SANTE PHYSIQUE	duree_sommeil	catégorielle	Durée moyenne de sommeil par nuit
	habitudes_alimentaires	Catégorielle	Qualité générale des habitudes alimentaires
SANTE MENTALE	pensees_suicidaires	Binaire	Présence pensées suicidaires
	stress_financier	Ordinale	Niveau de stress financier ressenti
	antecedents_familiaux_maladies mentales	Binaire	Présence d'antécédents familiaux de maladies mentales
	Depression	Binaire	Diagnostic de dépression

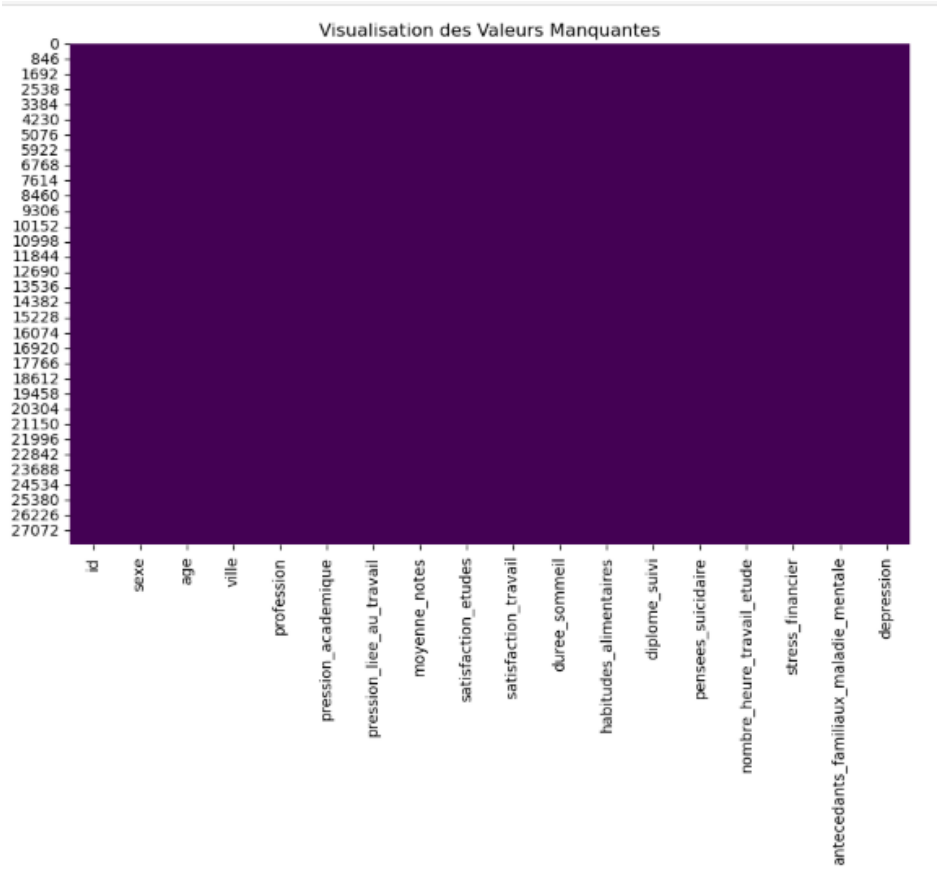
2- IMPORTATION DU JEU DE DONNEES

Dans le cadre d'une enquête sur la dépression 27901 étudiants ont été auditionnés dans l'optique de connaitre les facteurs entrainants et favorisants la dépression. Notre jeu de données contient 18 variables qui pourraient aider à comprendre les profondeurs de ce phénomène.

	id	sexe	age	ville	profession	pression_academique	pression_liee_au_travail	moyenne_notes	satisfaction_etudes	satisfaction_travail	duree_sommeil
0	2	Male	33.0	Visakhapatnam	Student	5.0	0.0	8.97	2.0	0.0	5-6 hours
1	8	Female	24.0	Bangalore	Student	2.0	0.0	5.90	5.0	0.0	5-6 hours
2	26	Male	31.0	Srinagar	Student	3.0	0.0	7.03	5.0	0.0	Less than 5 hours
3	30	Female	28.0	Varanasi	Student	3.0	0.0	5.59	2.0	0.0	7-8 hours
4	32	Female	25.0	Jaipur	Student	4.0	0.0	8.13	3.0	0.0	5-6 hours

3- TRAITEMENT DES VALEURS MANQUANTES

Les valeurs manquantes sont concentrées dans la colonne contenant la variable « Stress_financier » avec très peu de données manquantes, elles sont donc difficiles à voir.




```
Nombre de valeurs manquantes par colonne :
id                                      0
sexe                                    0
age                                     0
ville                                   0
profession                              0
pression_academique                     0
pression_liee_au_travail                0
moyenne_notes                           0
satisfaction_etudes                     0
 satisfaction_travail                   0
duree_sommeil                           0
habitudes_alimentaires                  0
diplome_suivi                           0
pensees_suicidaire                      0
nombre_heure_travail_etude              0
stress_financier                        3
antecedants_familiaux_maladie_mentale   0
depression                              0
dtype: int64
```

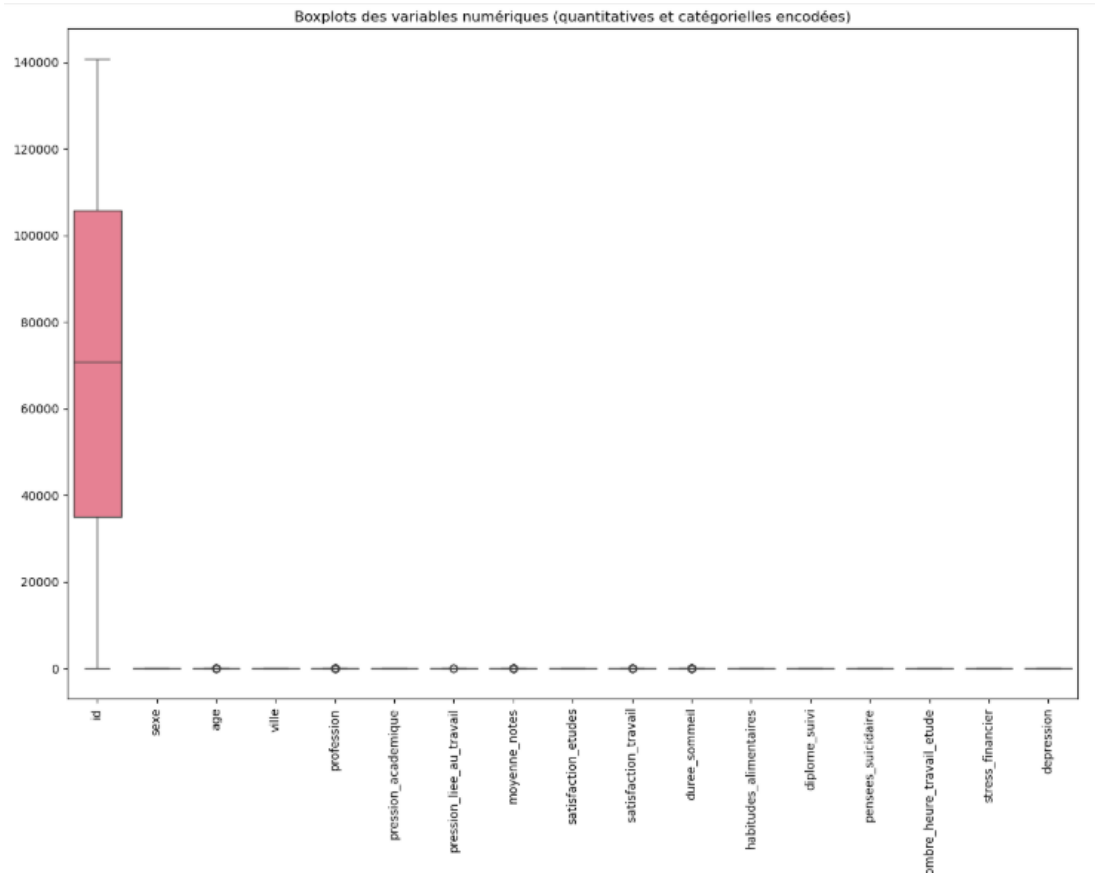
```
Pourcentage de valeurs manquantes par colonne :
id                                      0.000000
sexe                                    0.000000
age                                     0.000000
ville                                   0.000000
profession                              0.000000
pression_academique                     0.000000
pression_liee_au_travail                0.000000
moyenne_notes                           0.000000
satisfaction_etudes                     0.000000
 satisfaction_travail                   0.000000
duree_sommeil                           0.000000
habitudes_alimentaires                  0.000000
diplome_suivi                           0.000000
pensees_suicidaire                      0.000000
nombre_heure_travail_etude              0.000000
stress_financier                        0.010752
antecedants_familiaux_maladie_mentale   0.000000
depression                              0.000000
dtype: float64
```

On constate que contrairement à ce que la visualisation montre, il existe 3 valeurs manquantes dans la variables Stress_financier, elles représentent une proportion de 0,010752, ce qui correspond à environ 1,08%. Comme en général la proportion de valeurs manquantes inférieures au seuil de 5% sont considérées comme étant négligeables, elles ne nécessitent aucun traitement.

4-TRAITEMENT DES VALEURS ABERRANTES ET EXTREMES

Traiter les valeurs aberrantes permet d'assurer la qualité et la robustesse de notre analyse. Cela permet d'obtenir des résultats plus fiables, des modèles plus précis et des interprétations plus justes, tout en garantissant que notre analyse reflète fidèlement la réalité des données.

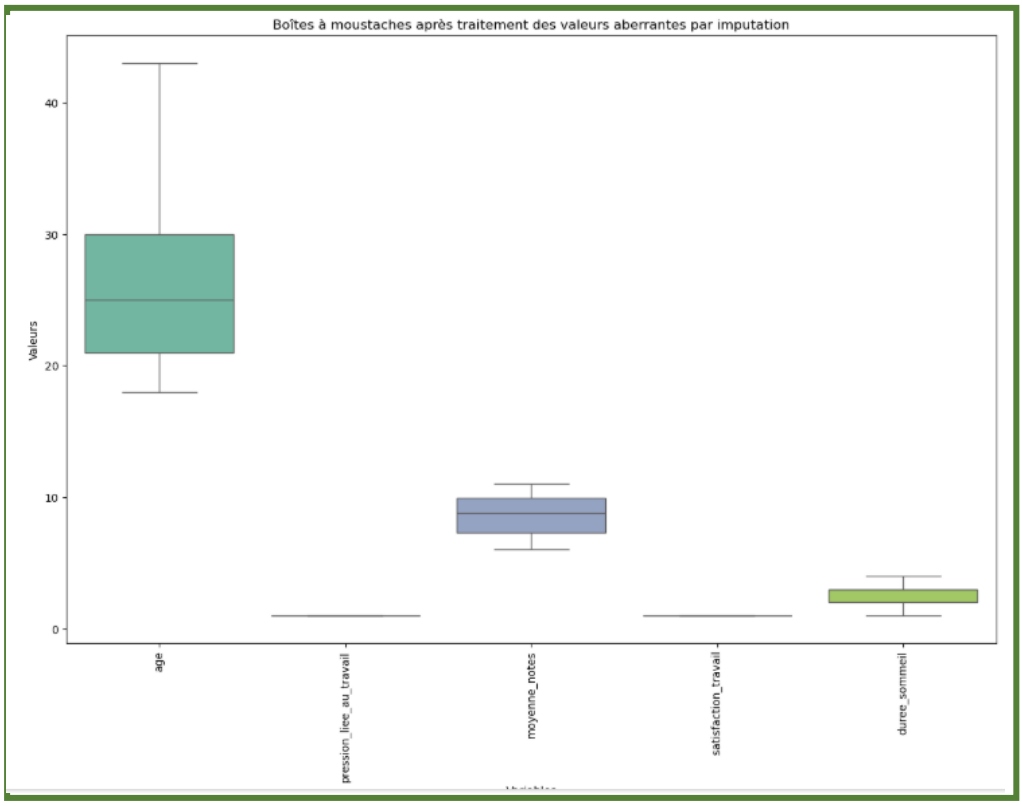
4-1.Visualisation avant traitement des valeurs aberrantes



```
Variables contenant des valeurs aberrantes :
- age : 12 valeurs aberrantes
- profession : 31 valeurs aberrantes
- pression_liee_au_travail : 3 valeurs aberrantes
- moyenne_notes : 9 valeurs aberrantes
-  satisfaction_travail : 8 valeurs aberrantes
- duree_sommeil : 18 valeurs aberrantes
```

4-2.Visualisation après traitement des valeurs aberrantes

Les variables age, profession, pression_liee_au_travail, moyenne_note, satisfaction_travail, duree_sommeil présentent un nombre significatif de valeurs aberrantes. Les transformations logarithmiques et racine carrée ont été testée pour réduire l'impact de ces variables extrêmes mais elles n'ont pas donné les résultats escomptés. La transformation logarithmique a amplifié les petites valeurs, créant de nouvelles valeurs aberrantes, en particulier pour la variable duree_someil, où le nombre de valeurs aberrantes est passé à 6183 après transformation. De même, la transformation racine carrée n'a pas suffisamment réduit l'asymétrie des distributions, laissant persister un nombre élevé de valeurs aberrantes. Ces méthodes ne sont pas adaptées à la nature des données, notamment pour la variable comme profession, qui est une variable catégorielle encodée, rendant les transformations mathématiques inappropriées. La méthode de l'imputation par la médiane est la méthode de traitement qui a permis de traiter efficacement les valeurs aberrantes.



```
Données après imputation des valeurs aberrantes :

Nombre de valeurs aberrantes après imputation :
age: 0 valeurs aberrantes
pression_liee_au_travail: 0 valeurs aberrantes
moyenne_notes: 0 valeurs aberrantes
 satisfaction_travail: 0 valeurs aberrantes
duree_sommeil: 0 valeurs aberrantes
```

5-TRAITEMENT DES DOUBLONS

Les doublons peuvent fausser les résultats en introduisant des biais, aucun doublon n'a été détecté, ce qui signifie que chaque observation est unique et contribue de manière équilibrée à l'analyse. Cette absence renforce la robustesse des résultats et permet une interprétation plus précise des données.

```
Nombre de doublons dans le jeu de données : 0
```

6-CATEGORISATION DES MODALITES

On catégorise les variables par niveaux de catégories pour faciliter la création des tableaux de contingence. Les variables comme ville, diplôme_suivi et profession sont de nombreuses modalités.

	taille_ville	secteur_activite	niveau_education	habitudes_alimentaires_categorie	duree_sommeil_categorie	categorie_age	sexe	pensees_suicidaire	antecedants
0	Grande Ville	autres	baccalauréat	Très sain	Hypersomnie	Adulte	Masculin	Oui	
1	Grande Ville	autres	baccalauréat	Moyen	Hypersomnie	Jeune adulte	Féminin	Non	
2	Grande Ville	autres	baccalauréat	Très sain	Insomnie	Adulte	Masculin	Non	
3	Grande Ville	autres	baccalauréat	Moyen	Sommeil normal	Jeune adulte	Féminin	Oui	

PARTIE 2 : STATISTIQUE DESCRIPTIVE

L'analyse descriptive vise à résumer et à décrire les principales caractéristiques des données. Elle permet de comprendre la distribution des variables, d'identifier des tendances générales et de préparer le terrain pour des analyses approfondies.

1-ANALYSE UNIVARIEE

	id	age	pression_academique	pression_liee_au_travail	moyenne_notes	satisfaction_etudes	satisfaction_travail	nombre_heure_travail_etude
Moyenne	70442.149421	25.822300	3.141214	0.000430	7.656104	2.943837	0.000681	7.156984
Médiane	70684.000000	25.000000	3.000000	0.000000	7.770000	3.000000	0.000000	8.000000
Écart-type	40641.175216	4.905687	1.381465	0.043992	1.470707	1.361148	0.044394	3.707642
IQR	70779.000000	9.000000	2.000000	0.000000	2.630000	2.000000	0.000000	6.000000
Minimum	2.000000	18.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Maximum	140699.000000	59.000000	5.000000	5.000000	10.000000	5.000000	4.000000	12.000000

2- INTERPRETATION

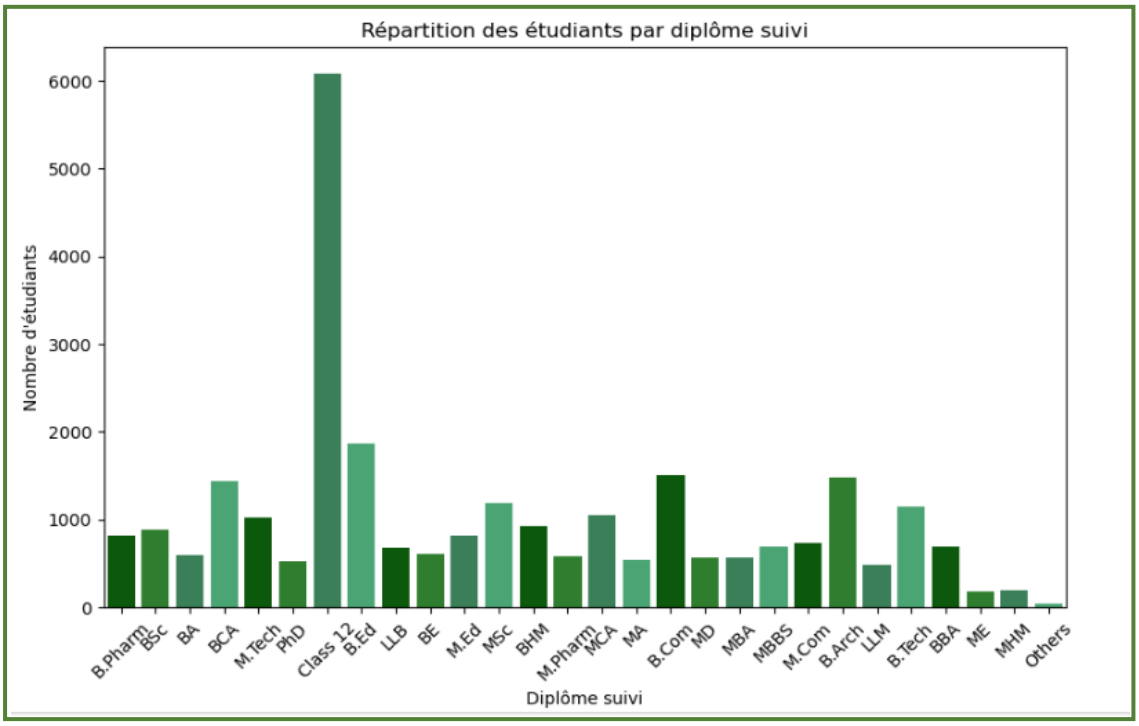
Les variables pression académique, pression liée au travail, stress financier semblent mesurer le niveau de stress ressenti par les étudiants. Des valeurs élevées pourraient indiquer un risque accru de dépression.
satisfaction_etudes et satisfaction _travail mesurent le niveau de satisfaction des étudiants vis-à-vis de leurs études et de leur travail. Une faible satisfaction pourrait être un facteur de risque de dépression.
Les variables durée_sommeil,et habitudes_alimentaires reflètent le mode de vie des étudiants et pourraient avoir un impact sur leur santé mentale.
La variable antécédente familiaux de maladies mentales pourrait fournir des informations sur les facteurs de risque génétiques ou environnementaux.

3- VISUALISATIONS

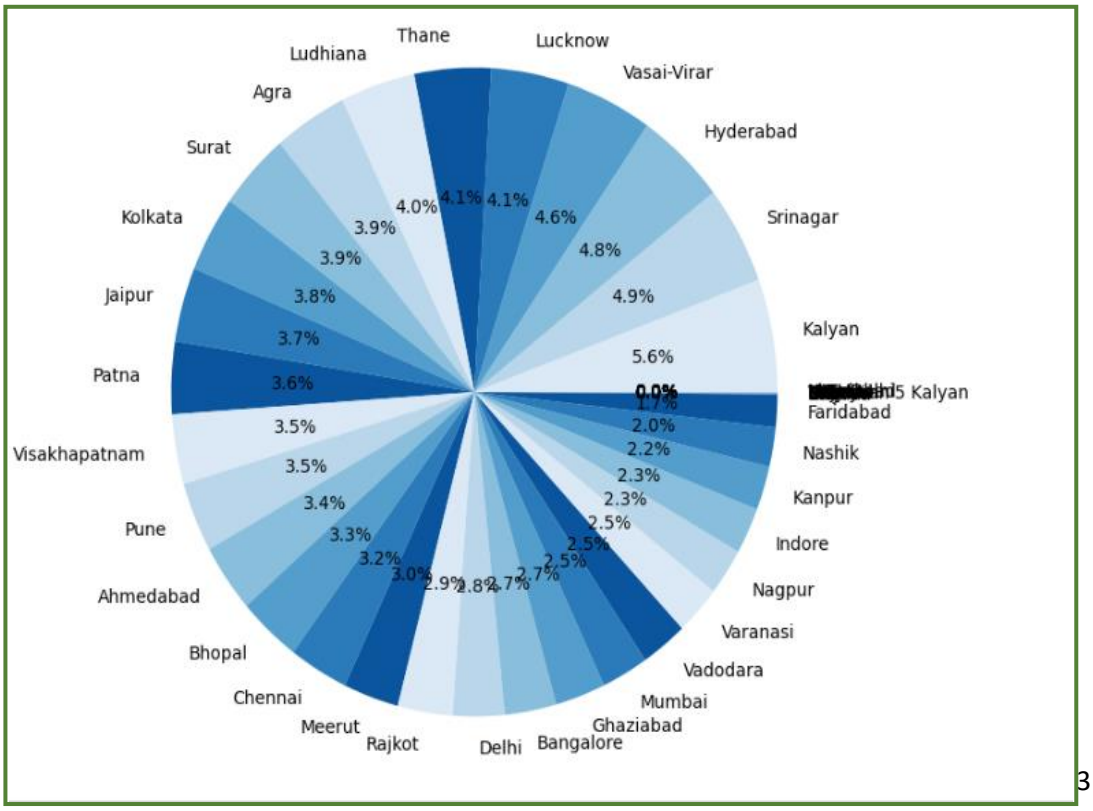
3-1.Diagramme en bande de la variable diplome_suivi

Ce graphique en barres verticales représente la distribution des étudiants en fonction du diplôme qu'ils suivent. Chaque barre correspond à un type de diplôme, et la hauteur de la barre indique le nombre d'étudiants inscrits dans ce diplôme. Le graphique nous indique clairement que la majorité des étudiants sont inscrits en

"Class 12"**.** Les modalités B.Ed (Baccalauréat en Éducation) montre un nombre significatif d'étudiants se préparent à devenir enseignants, B.Com (Baccalauréat en Commerce) et B.Arch (Baccalauréat en Architecture) attirent un nombre conséquent d'étudiants.



3-2.Diagrammes en Camembert



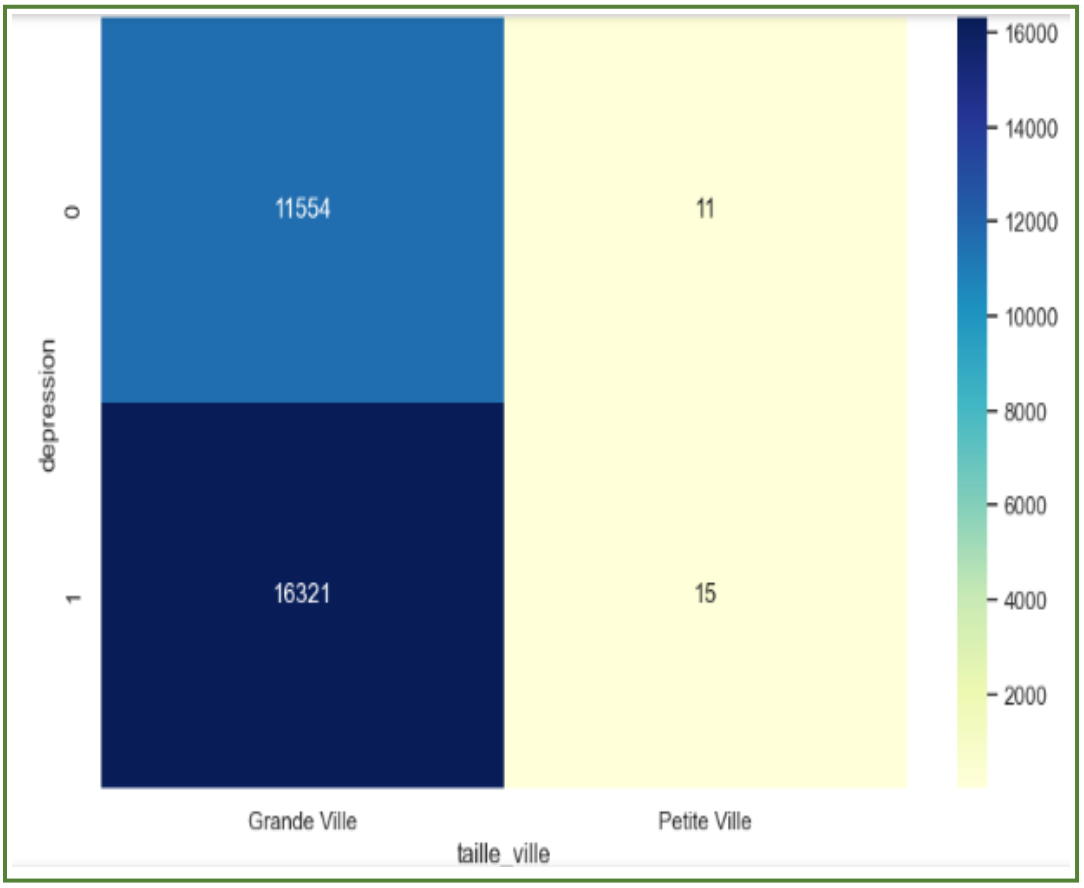
Les résultats suggèrent une prévalence plus élevée de symptômes dépressifs dans les grandes métropoles telles que Mumbai, Delhi et Bangalore. Ces villes concentrent une proportion significativement plus importante de la population étudiée présentant des signes de dépression.

4- STATISTIQUE BIVARIEE

Croiser la variable de dépression avec les autres variables catégorielles permet de voir quels facteurs sont liés à la dépression chez les étudiants. Ce facteur permet d'identifier les tendances pour mieux comprendre ce qui influence la santé mentale et proposer des solutions adaptées.

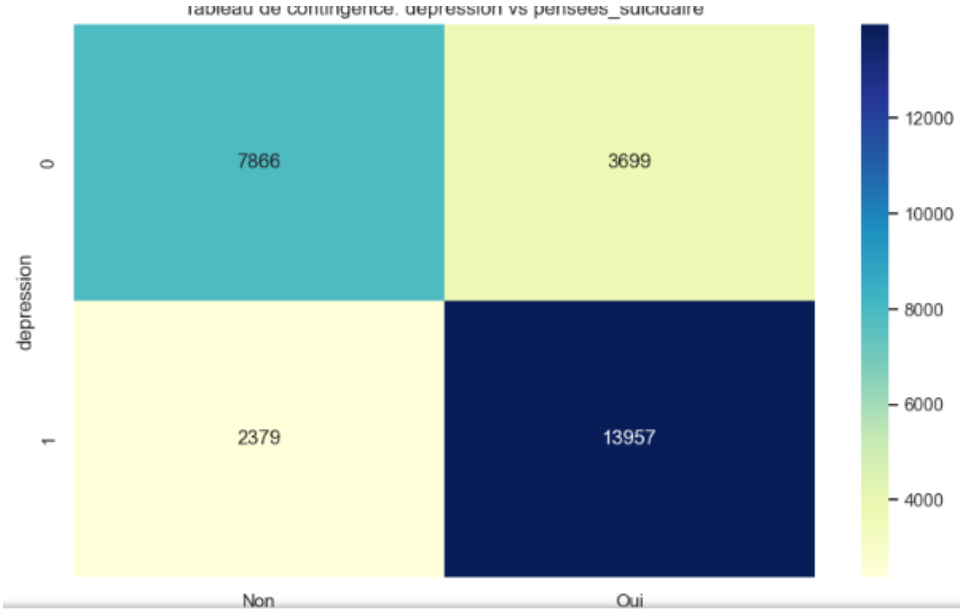
4-1. Croisement entre la variable ville et dépression

Les modalités de certaines variables comme ville, profession et diplome_suivi étant beaucoup trop nombreuses, nous avons défini chaque modalité par niveaux. Les niveaux de catégorisation se font pour la variable ville en fonction de la superficie des villes, nous avons donc deux niveaux de catégorisation (petites et grandes villes)



Le nombre de personnes déclarant une dépression est bien plus élevé dans les grandes villes que dans les petites. Les cases correspondant aux grandes villes et à la dépression sont nettement plus sombres, indiquant un nombre significativement plus important de cas.
À l'inverse, les petites villes semblent moins touchées par la dépression, avec des chiffres nettement plus faibles.

4-2. Croisement entre la variable pensees_suicidaires et dépression

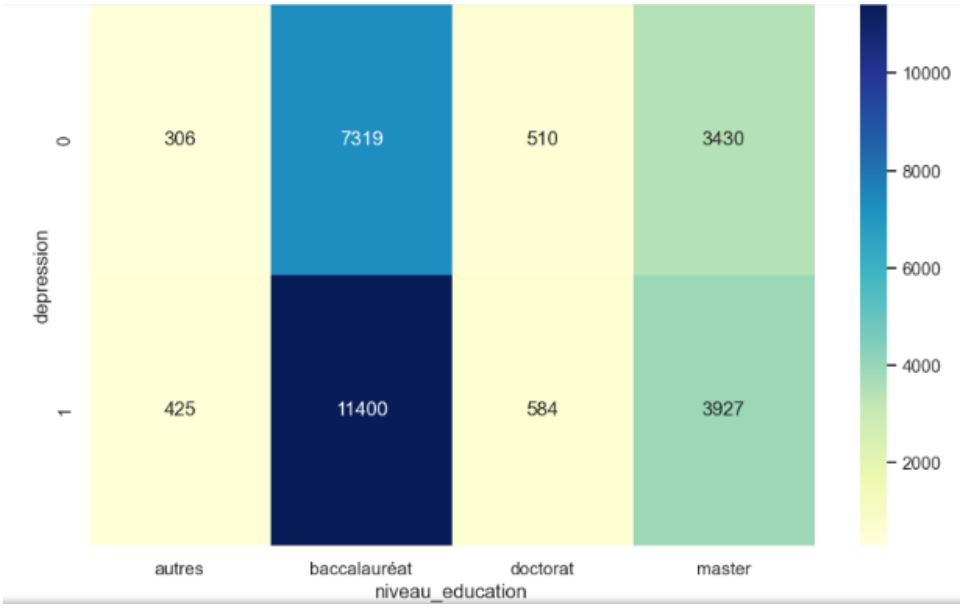


Le tableau révèle une relation très claire entre la dépression et les pensées suicidaires :

On observe un nombre significativement plus élevé d'individus qui présentent à la fois une dépression et des pensées suicidaires (coin supérieur droit).

Les autres cellules, représentant les individus sans dépression et sans pensées suicidaires, ou bien avec dépression mais sans pensées suicidaires, ont des effectifs nettement plus faibles.

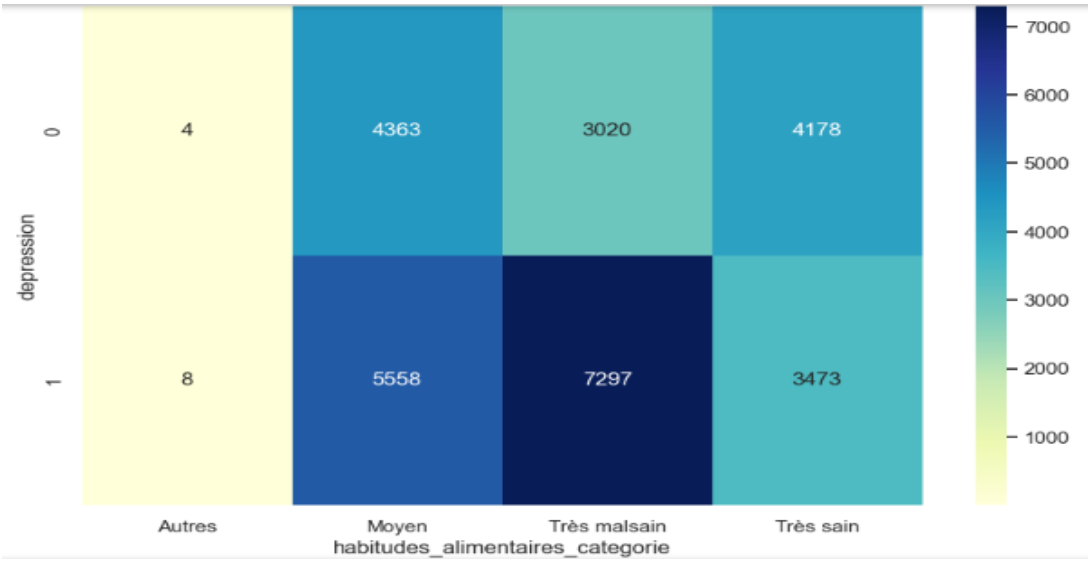
4-3. Croisement entre la variable diplôme_suivi et dépression



Le tableau suggère une relation entre le niveau d'éducation et la dépression.

La catégorie "baccalauréat" présente le plus grand nombre de cas de dépression. Cela pourrait indiquer que les personnes ayant un niveau d'études baccalauréat sont plus susceptibles de souffrir de dépression par rapport aux autres groupes. Les autres catégories (autres, doctorat, master) présentent des nombres de cas de dépression moins élevés. Plusieurs hypothèses peuvent expliquer ces résultats : Les personnes ayant un baccalauréat peuvent être confrontées à des exigences professionnelles plus élevées, à une plus grande compétition, ou à des attentes sociales plus importantes, ce qui pourrait augmenter leur risque de dépression. Le niveau d'éducation peut être lié à d'autres facteurs socio-économiques comme le revenu, le statut social, ou les conditions de travail, qui peuvent tous influencer la santé mentale

4-4. Croisement entre la variable habitudes_alimentaire et dépression



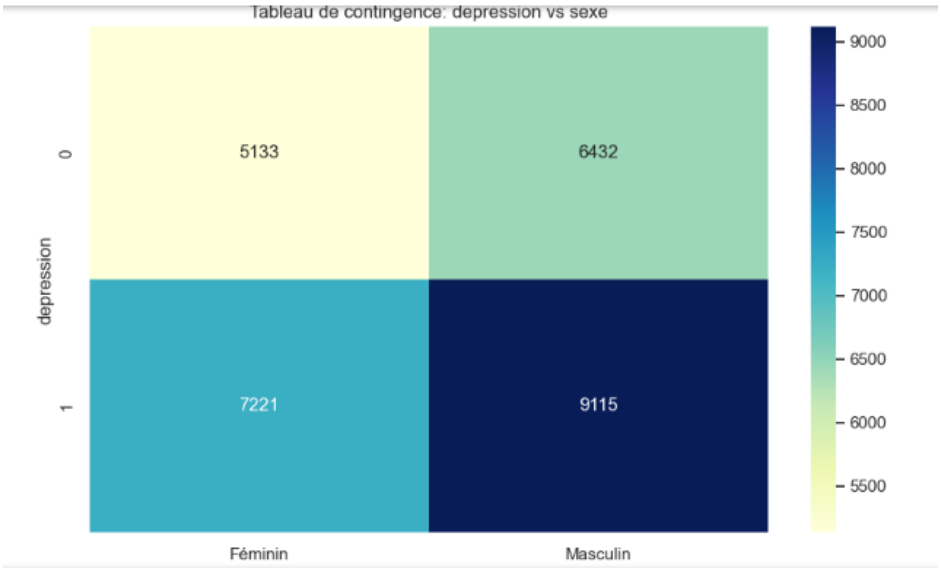
Le tableau suggère une relation entre les habitudes alimentaires et la dépression.

La catégorie "très malsaines" présente le plus grand nombre de cas de dépression. Cela pourrait indiquer que les personnes ayant des habitudes alimentaires très malsaines sont plus susceptibles de souffrir de dépression par rapport aux autres groupes.

À l'inverse, la catégorie "très saines" présente un nombre légèrement inférieur de cas de dépression, suggérant une possible association positive entre des habitudes alimentaires saines et une meilleure santé mentale.

La nutrition joue un rôle important dans la régulation de l'humeur et des neurotransmetteurs. Une alimentation déséquilibrée peut contribuer à l'apparition de troubles dépressifs.
Les personnes ayant des habitudes alimentaires malsaines sont souvent associées à d'autres comportements à risque pour la santé, comme la sédentarité, la consommation d'alcool ou de tabac, qui peuvent également favoriser la dépression.
Les habitudes alimentaires peuvent être liées à d'autres facteurs socio-économiques comme le niveau de revenu, le niveau d'éducation, le statut social, qui peuvent tous influencer la santé mentale.

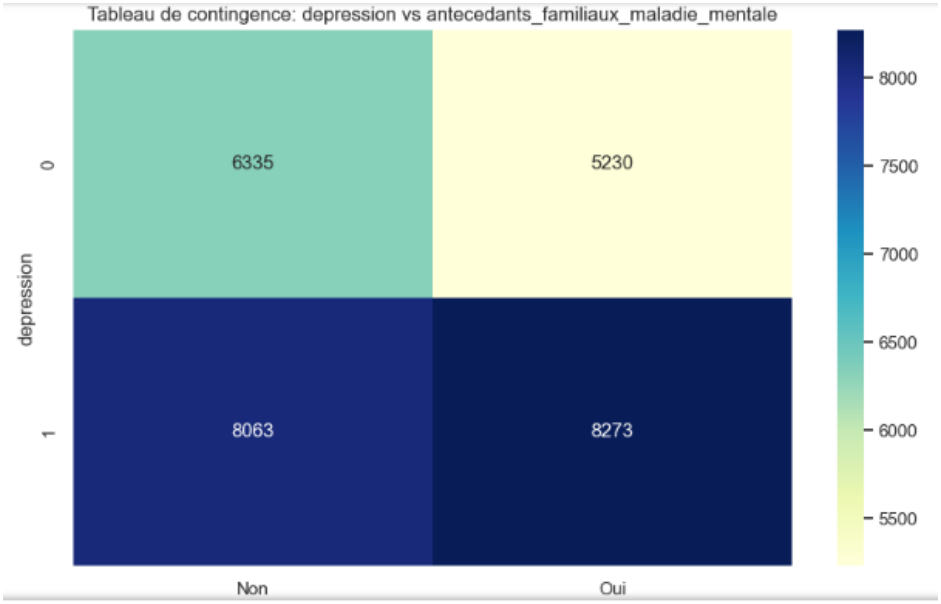
4-5. Croisement entre la variable sexe et dépression



Le tableau suggère une légère différence entre les sexes en termes de prévalence de la dépression. Les femmes semblent légèrement plus touchées par la dépression que les hommes.

Les femmes sont souvent confrontées à des rôles sociaux et des attentes spécifiques qui peuvent augmenter leur vulnérabilité à la dépression.
Les différences hormonales entre les hommes et les femmes peuvent jouer un rôle dans la susceptibilité à la dépression.
Il est possible que les femmes soient plus à l'aise pour exprimer leurs émotions et chercher de l'aide, ce qui pourrait expliquer une plus grande proportion de femmes déclarant une dépression.

4-6. Croisement entre la variable antecedants_familiaux_maladie_mentale et dépression



Le nombre de personnes ayant des antécédents familiaux de maladie mentale et souffrant de dépression est plus élevé comparé au nombre de personnes sans antécédents familiaux et souffrant de dépression. Cela suggère que les personnes ayant des antécédents familiaux sont plus susceptibles de développer une dépression.

Les antécédents familiaux de maladie mentale peuvent suggérer une prédisposition génétique à développer des troubles mentaux, y compris la dépression.

Les membres d'une même famille partagent souvent un environnement similaire (éducation, style de vie, etc.) qui peut influencer le développement de troubles mentaux.

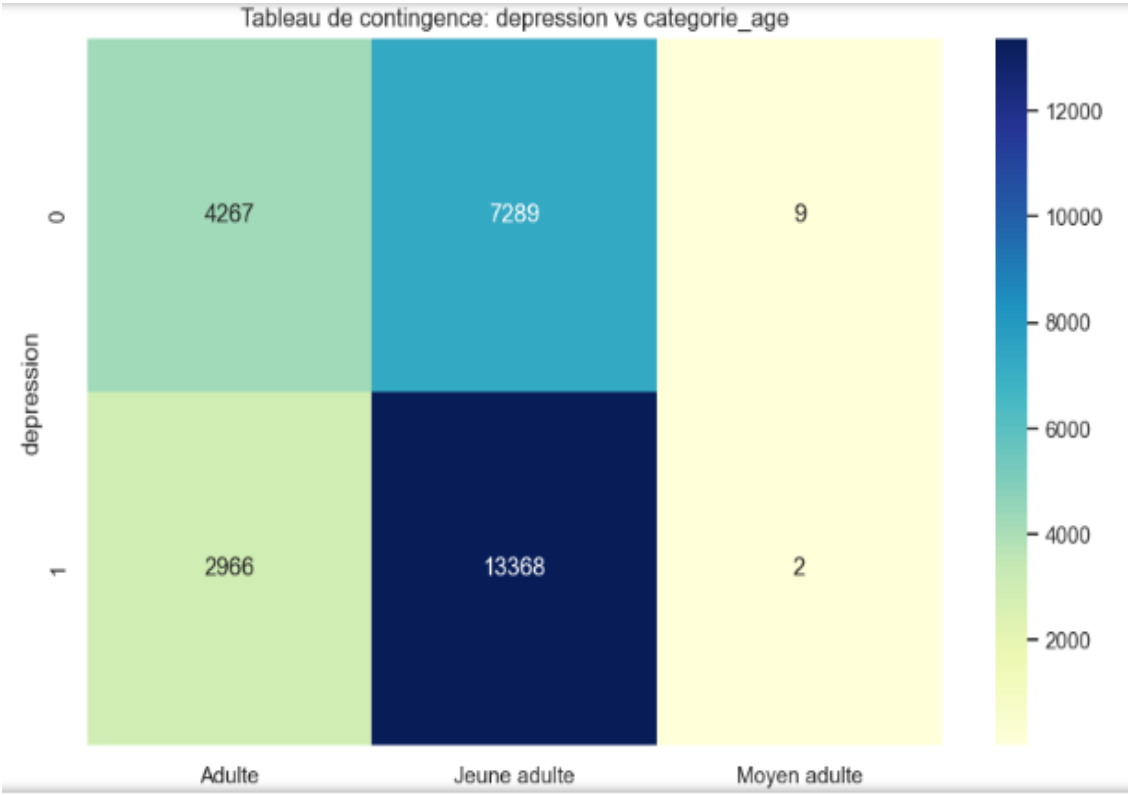
4-7. Croisement entre la variable duree_sommeil et dépression

La catégorie "insomnie" présente le plus grand nombre de cas de dépression. Cela pourrait indiquer que les personnes souffrant d'insomnie sont plus susceptibles de développer une dépression.
La catégorie "hypersomnie" présente également un nombre significatif de cas de dépression, bien qu'un peu moins élevé que pour l'insomnie.

La catégorie "sommeil normal" présente le nombre le plus faible de cas de dépression, suggérant une association positive entre un sommeil de qualité et une meilleure santé mentale.

Les troubles du sommeil, qu'il s'agisse d'insomnie ou d'hypersomnie, sont souvent associés à des troubles de l'humeur comme la dépression. Les difficultés à s'endormir, à maintenir le sommeil ou à se réveiller trop tôt peuvent contribuer à l'apparition de symptômes dépressifs.

4-8. Croisement entre la variable age et dépression



La catégorie "jeune adulte" présente le plus grand nombre de cas de dépression. Cela pourrait indiquer que les jeunes adultes sont plus susceptibles de développer une dépression par rapport aux autres groupes d'âge.

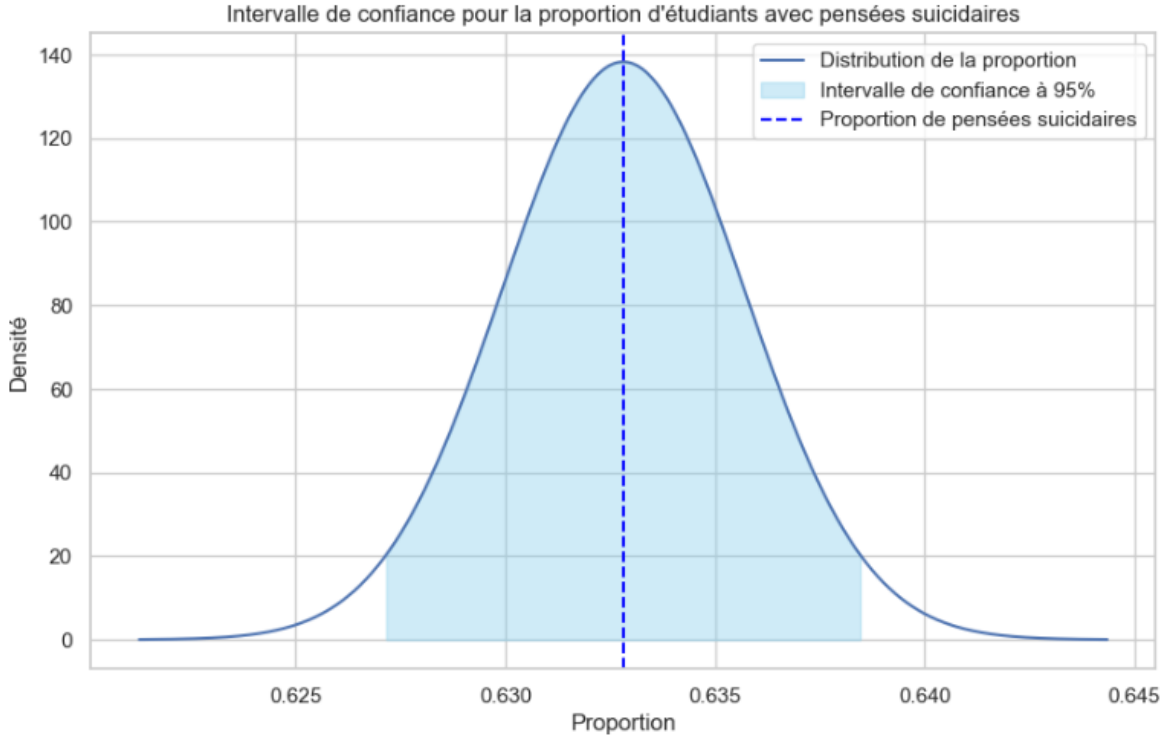
Les catégories "adulte" et "moyen adulte" présentent un nombre de cas de dépression significativement inférieur par rapport à la catégorie "jeune adulte".
Les jeunes adultes traversent souvent des périodes de transition importantes (études, entrée dans la vie professionnelle, construction d'un projet de vie) qui peuvent être sources de stress et favoriser l'apparition de troubles dépressifs.
Les jeunes adultes peuvent être plus exposés à certains facteurs de risque pour la santé mentale, tels que la pression sociale, les difficultés financières, ou des événements de vie stressants.
Les jeunes adultes sont plus susceptibles de consommer des substances psychoactives, ce qui peut augmenter le risque de dépression.
Les jeunes adultes peuvent être moins enclins à consulter un professionnel de la santé pour des problèmes de santé mentale, ce qui pourrait sous-estimer la prévalence de la dépression dans ce groupe d'âge.

PARTIE 3 : ANALYSE DES OBJECTIFS DU PROJET

Cette partie a pour objectif d'analyser les données collectées afin de répondre aux questions de recherche spécifiques. Nous commencerons par estimer la prévalence des pensées suicidaires au sein de la population étudiante en calculant un intervalle de confiance. Ensuite, nous comparerons les étudiants dépressifs et non-dépressifs en termes de charge de travail, de stress financier et de satisfaction. Des tests statistiques appropriés seront utilisés pour déterminer si ces différences sont significatives. Enfin, nous explorerons les liens entre la dépression, les habitudes alimentaires et la durée du sommeil. Les résultats de cette analyse nous permettront de mieux comprendre les facteurs associés à la dépression chez les étudiants et d'identifier les groupes les plus vulnérables.

1- INTERVALLE DE CONFIANCE DES ETUDIANTS AYANT DES PENSEES SUICIDAIRES (Question 1)

```
Proportion d'étudiants avec pensées suicidaires: 0.6328
Intervalle de confiance pour les pensées suicidaires : (0.6271527143195701, 0.6384650054754193)
```



En moyenne, 63,28% des étudiants de votre échantillon ont déjà eu des pensées suicidaires.

Nous sommes très confiants (à 95%) que la proportion réelle d'étudiants ayant déjà eu des pensées suicidaires dans la population totale se situe entre 62,72% et 63,85%

Le graphique que vous présentez représente une distribution de probabilité. La courbe en bleu représente la distribution de toutes les proportions possibles d'étudiants ayant des pensées suicidaires, si l'on répétait l'étude un grand nombre de fois. La zone ombrée en bleu clair représente l'intervalle de confiance à 95%. Les valeurs numériques indiquent que l'intervalle de confiance pour la proportion d'étudiants ayant déjà eu des pensées suicidaires se situe entre 0,6271527143195701 et 0,6384650054754193. Cela signifie qu'on est à 95% sûr que la vraie proportion d'étudiants ayant eu des pensées suicidaires dans la population étudiée se situe entre ces deux valeurs.

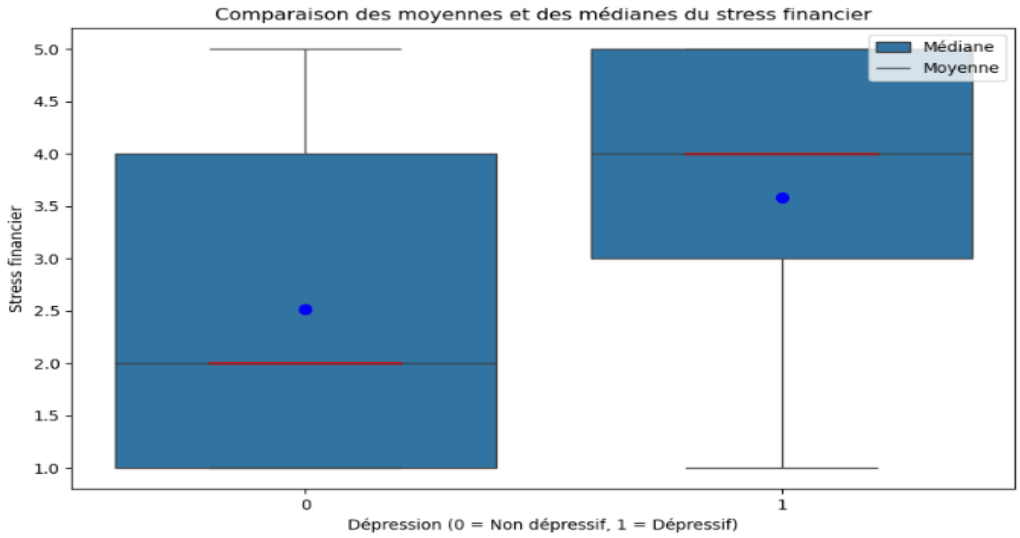
2- INDENTIFICATION DU TEST STATISTIQUE ADAPTE (Question 2)

Dans notre étude, nous comparons deux groupes d'étudiants distincts et indépendants : ceux qui éprouvent un stress financier et sont également dépressifs, et ceux qui, malgré un stress financier, ne présentent pas de symptômes dépressifs.

2-1.Test d'ajustement a une loi

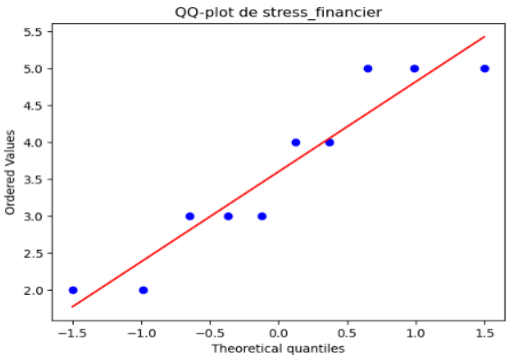
Les tests d'ajustement ont montré que les variables stress_financier et nombre_heure_travail_etude suit une distribution normale, contrairement à la variable dépression qui est non normale, elle est binaire donc elle suit une loi binomiale. Or, lorsque les données ne suivent pas au moins une distribution normale, le test de Mann-Whitney-Wilcoxon est privilégié. Ce test non paramétrique permet de comparer les médianes, sans faire d'hypothèse sur la distribution des données.

2-2.Visualisation de moyenne et de médiane des populations

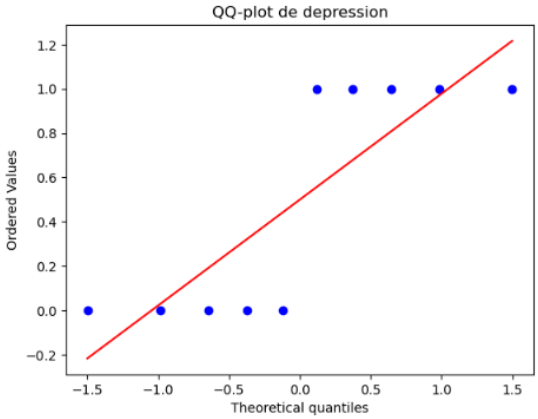


La médiane du stress financier est nettement plus élevée chez les personnes dépressives (1) que chez les personnes non-dépressives (0). Cela signifie qu'en moyenne, les personnes dépressives ressentent un stress financier plus important.
La boîte représentant les personnes dépressives est plus large, ce qui indique une plus grande dispersion des données. Cela signifie que le niveau de stress financier varie davantage au sein du groupe des personnes dépressives.

2-3.Visualisation de test ajustement stress_financier et depression



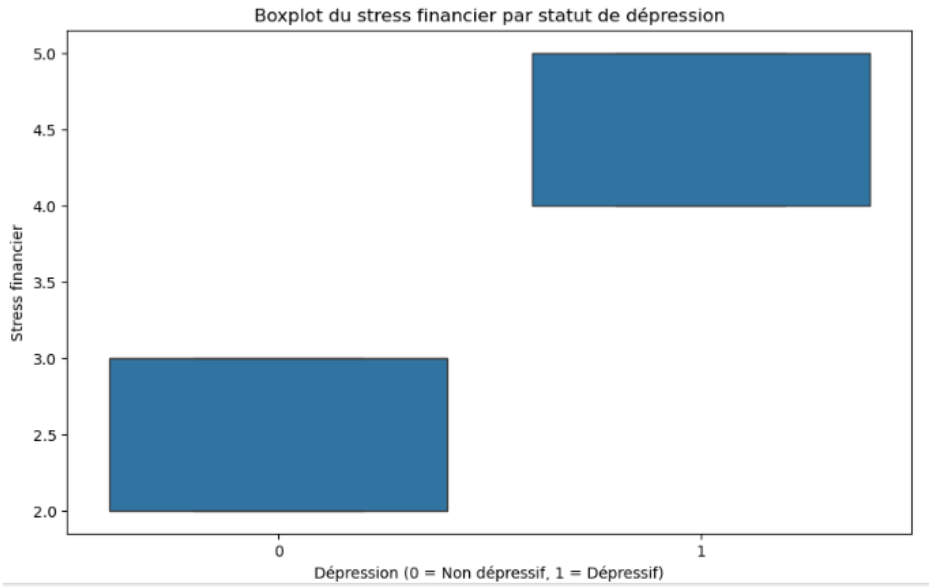

```
Statistique de Shapiro-Wilk (stress_financier) : 0.878029749889342
Valeur p (stress_financier) : 0.1238586271644549
L'hypothèse nulle de normalité pour stress_financier ne peut pas être rejetée (données norm
```




```
Statistique de Shapiro-Wilk (depression) : 0.6552710244620128
Valeur p (depression) : 0.0002539627375607894
L'hypothèse nulle de normalité pour depression est rejetée (données non normalement distribuées).
```

Test de significativité

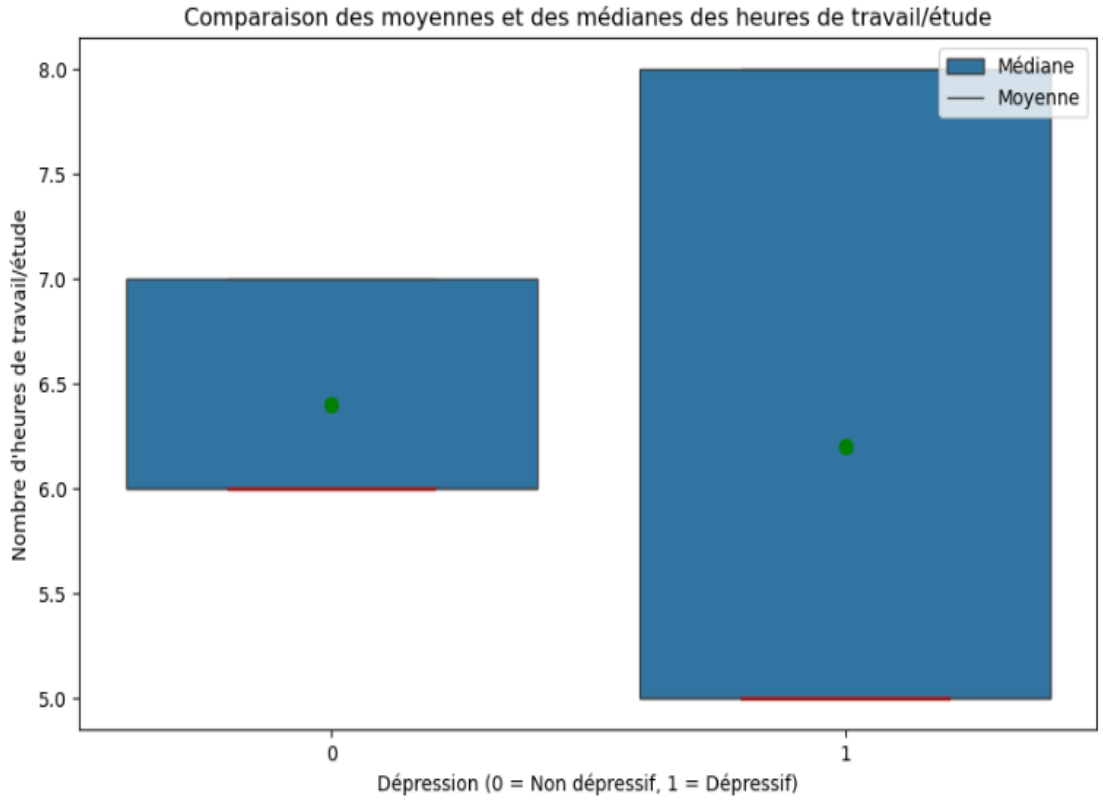
```
Statistique de Wilcoxon (U) : 25.0
Valeur p : 0.009700785068229596
Différence significative entre les groupes (hypothèse nulle rejetée).
```



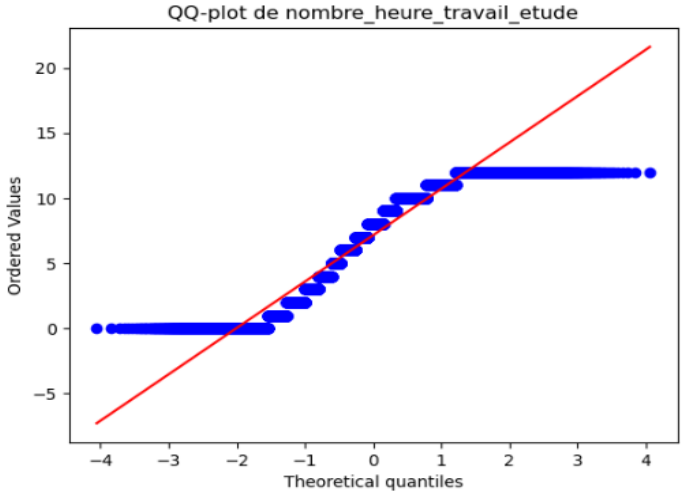
Ces résultats suggèrent un lien étroit entre la dépression et le stress financier. Les personnes souffrant de dépression semblent éprouver un niveau de stress financier significativement plus élevé que celles qui ne sont pas dépressives.

2-4.Visualisation de moyenne et de médiane des populations

La médiane du nombre d'heures travaillées/étudiées est nettement plus élevée chez les personnes non-dépressives (0) que chez les personnes dépressives (1). Cela signifie qu'en moyenne, les personnes non - dépressives consacrent plus de temps au travail ou aux études.
La boîte représentant les personnes non-dépressives est légèrement plus large, ce qui indique une plus grande dispersion des données. Cela signifie que le nombre d'heures travaillées/étudiées varie davantage au sein du groupe des personnes non-dépressives.
On observe quelques valeurs aberrantes dans les deux groupes, notamment une personne non-dépressive qui travaille ou étudie beaucoup plus que les autres.



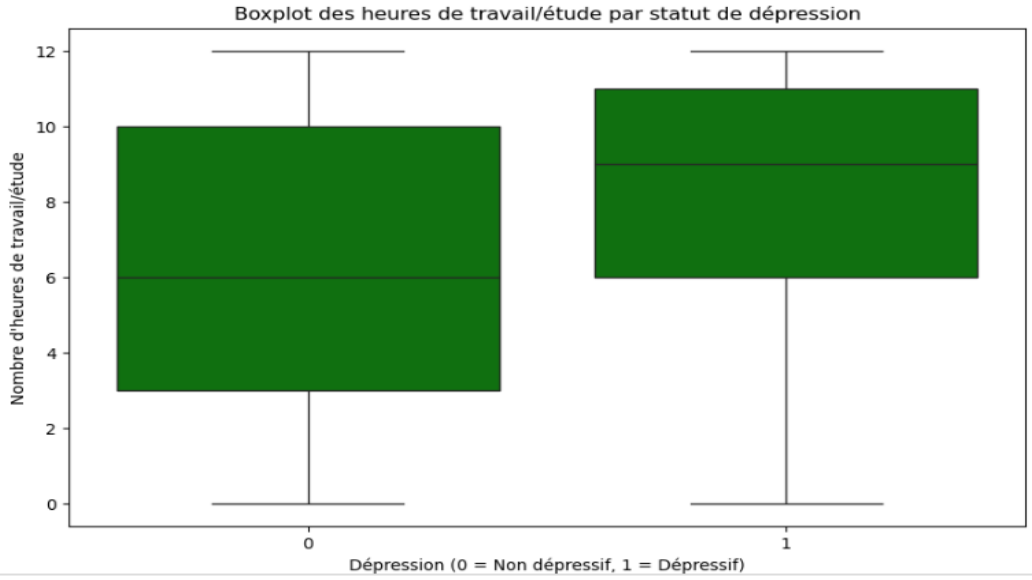
2-5. Visualisation test de conformité nombre_heure_travail_etude et depression




```
Statistique de Shapiro-Wilk (nombre_heure_travail_etude) : 0.9213483468791226
Valeur p (nombre_heure_travail_etude) : 1.5056890316122357e-78
L'hypothèse nulle de normalité pour nombre_heure_travail_etude est rejetée (données non normalement distribuées).
```

2-6. Test de significativité

```
Statistique de Wilcoxon (U) : 116578577.0
Valeur p : 2.55320335547491e-246
Différence significative entre les groupes (hypothèse nulle rejetée).
```



Les personnes dépressives consacrent en moyenne autant de temps au travail ou aux études que les personnes non-dépressives.
La dépression n'impacte pas de manière significative la quantité de temps passée à travailler ou à étudier

2-6. TABLEAU DE CONTINGENCES

```
Tableau de contingence :
stress_financier            1.0                                                \
nombre_heure_travail_etude 0.0  1.0  2.0  3.0  4.0  5.0  6.0  7.0  8.0  9.0
depression
0                           337  204  272  227  250  187  272  255  246  239
1                            51   47   82   71   82   70  129  131  142  121

stress_financier            ...  5.0                                          \
nombre_heure_travail_etude  ... 3.0  4.0  5.0  6.0  7.0  8.0  9.0   10.0 11.0
depression                  ...
0                           ...   96   89   67   97   88  109   86   120  105
1                           ...  246  259  248  411  372  510  400  1031  685

stress_financier
nombre_heure_travail_etude 12.0
depression
0                           108
1                           705
```

2-6. ESTIMATION DES MEDIANES

```
Médianes du stress financier :
Étudiants dépressifs : 4.0
Étudiants non dépressifs : 2.0

Médianes des heures de travail/étude :
Étudiants dépressifs : 9.0
Étudiants non dépressifs : 6.0
```

Stress financier :

- Étudiants dépressifs: La médiane est de 4.0. Cela signifie que 50% des étudiants dépressifs ont un niveau de stress financier supérieur ou égal à 4 (sur une échelle que vous n'avez pas précisée).

- Étudiants non-dépressifs: La médiane est de 2.0. Cela signifie que 50% des étudiants non-dépressifs ont un niveau de stress financier supérieur ou égal à 2.

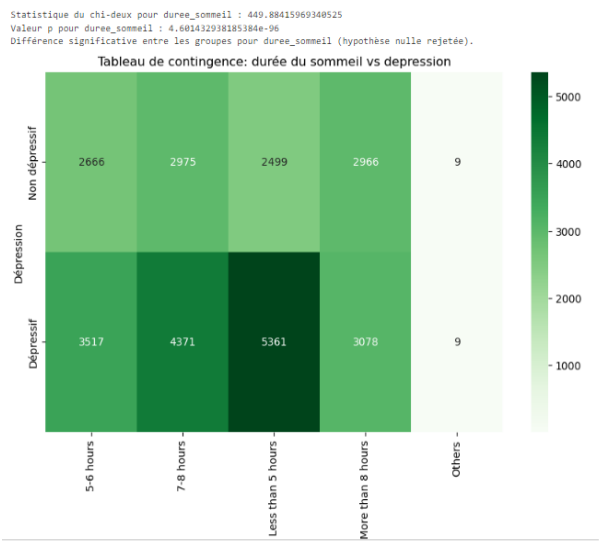
Nombre d'heures de travail/étude**:**

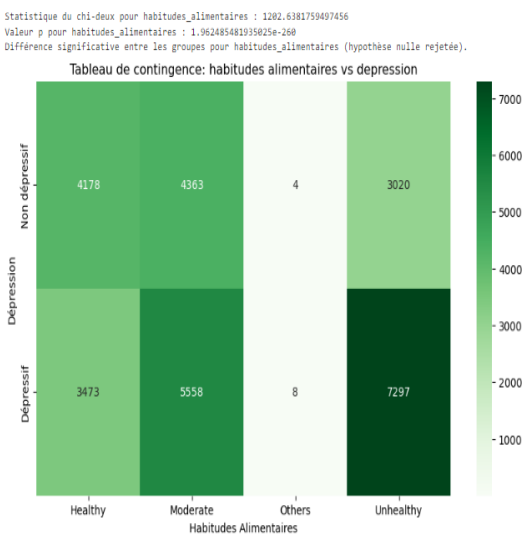
- Étudiants dépressifs: La médiane est de 9.0. Cela signifie que 50% des étudiants dépressifs travaillent ou étudient 9 heures ou plus par semaine.
- Étudiants non-dépressifs**:** La médiane est de 6.0. Cela signifie que 50% des étudiants non-dépressifs travaillent ou étudient 6 heures ou plus par semaine.

3- TEST D'INDEPENDANCE (Question 4)

Les variables habitudes_alimentaires, duree_sommeil et depression sont des variables qualitatives. Tester l'indépendance entre les variables habitudes alimentaires et depression puis ensuite les variables duree_sommeil et depression necesite un test d'independance entre deux variables qualitatives. Le test de Fisher est particulièrement adapté aux tableaux de contingence et ne nécessite pas d'hypothèses strictes sur la taille de l'échantillon. Cependant, étant donné la taille importante de votre jeu de données (27901 lignes), le test du khi-deux est plus adapté.

3-1. Visualisations tests d'indépendances





Les tests de khi-deux permettent de dire que les personnes dépressives ont une durée de sommeil significativement différente des personnes non dépressives.

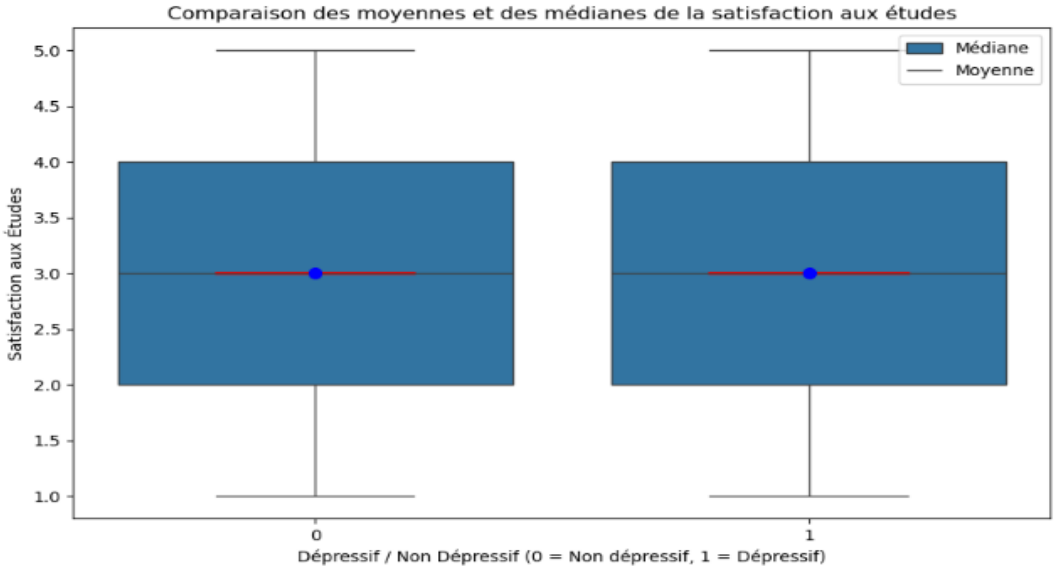
Les personnes dépressives ont des habitudes alimentaires significativement différentes des personnes non dépressives.

Une faible valeur de p (typiquement inférieure à 0,05), ce qui signifie qu'il est extrêmement improbable d'observer une telle association entre la dépression et les habitudes alimentaires par hasard. On rejette l'hypothèse nulle (qui stipule qu'il n'y a pas de lien entre les deux variables) et on conclut qu'il existe bien une relation significative.

4- TEST DE COMPARAISON DE MEDIANES (Question 3)

Les données de la variable satisfaction_etude ne suivent pas la loi normale et les variables depressifs§non depressifs sont des donnés binaires elles suivent la loi-binomiale seul la distribution de satisfaction_etude qui suit une loi non normale permet de dire que le test de wicolxon est adapté.

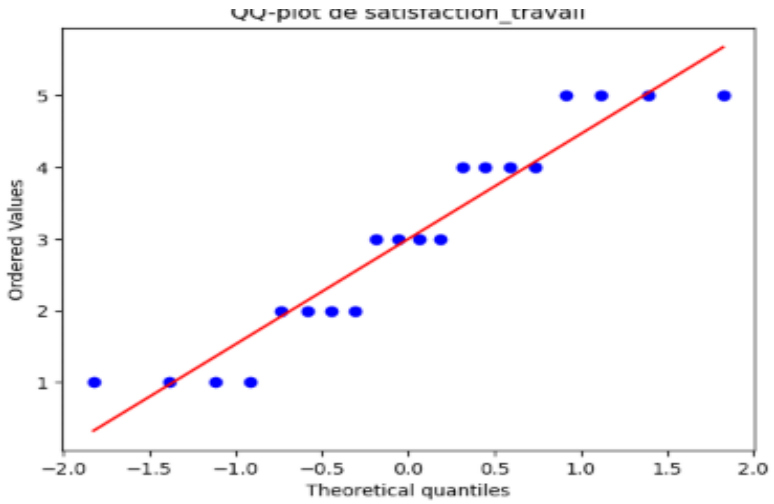
```
Tableau de contingence :
satisfaction_etude       1  2  3  4  5
depressif_non_depressif
0                        2  2  2  2  2
1                        2  2  2  2  2
```




```
Test de Wilcoxon : Statistique = 19.0, p-value = 0.431640625
```

Étant donné que la p-valeur (0.4316) est supérieure au seuil de signification classique de 0.05, nous ne pouvons pas rejeter l'hypothèse nulle. Cela signifie que nous n'avons pas suffisamment de preuves pour affirmer qu'il existe une différence significative entre la satisfaction aux études des étudiants dépressifs et non-dépressifs.

4-1. Test de comparaison de moyennes




```
Statistique de Shapiro-Wilk (satisfaction_travail) : 0.8958835123690987
Valeur p (satisfaction_travail) : 0.034548326289013 79
L'hypothèse nulle de normalité pour satisfaction_travail est rejetée (données non normalement distribuées).
Test de Kruskal-Wallis : Statistique = 18.999999999999993, p-value = 0.45683612559196274
```

La distribution de la satisfaction au travail n'est pas parfaitement normale mais s'en approche.
Il n'y a pas de différence significative entre les médianes des différents groupes

5- CONCLUSION GENERALE

Les résultats obtenus suggèrent que les étudiants dépressifs sont plus susceptibles de faire face à des défis tels que le stress financier, une surcharge de travail et des difficultés à maintenir des habitudes de vie saines.

6- RECOMANDATIONS

Pour améliorer la situation des étudiants dépressifs, nous recommandons de:

- Mettre en place des mesures de soutien financier.
- Proposer des outils de gestion du temps et du stress.
- Favoriser un accès plus facile aux soins de santé mentale.
- Créer un environnement universitaire plus bienveillant.

CODE SOURCE PYTHON

```
# Importer la bibliothèque
import pandas as pd
import os
# Chemin du fichier CSV
df = pd.read_csv('C:/Users/hp/Downloads/KEX/Student_Depression.csv', sep=',')
# Affichez les premières lignes du jeu de données
df.head()

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Visualisation des valeurs manquantes
plt.figure(figsize=(10, 6))
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.title('Visualisation des Valeurs Manquantes')
plt.show()

# Pourcentage de valeurs manquantes par colonne
missing_percentage = df.isnull().mean() * 100
print("Pourcentage de valeurs manquantes par colonne :")
print(missing_percentage)

# Compter le nombre de doublons
nombre_doublons = df.duplicated().sum()

# Afficher le nombre de doublons
print(f"Nombre de doublons dans le jeu de données : {nombre_doublons}")

#visualisation valeurs abberantes import seaborn as sns
import matplotlib.pyplot as plt
# Sélectionner les colonnes numériques (quantitatives et catégorielles encodées)
numeric_columns = df.select_dtypes(include=['float64', 'int64']).columns
print("\nColonnes numériques sélectionnées :")
print(numeric_columns)
# Tracer les boxplots pour chaque colonne numérique sur un seul graphique
plt.figure(figsize=(15, 10))
sns.boxplot(data=df[numeric_columns])
plt.xticks(rotation=90)  # Rotation des étiquettes pour qu'elles soient lisibles
plt.title("Boxplots des variables numériques (quantitatives et catégorielles
encodées)")
plt.show()

#Traitement valeurs aberrantes
import seaborn as sns
import matplotlib.pyplot as plt
# Sélectionner les colonnes numériques (quantitatives et catégorielles encodées)
numeric_columns = df.select_dtypes(include=['float64', 'int64']).columns
print("\nColonnes numériques sélectionnées :")
print(numeric_columns)
# Tracer les boxplots pour chaque colonne numérique sur un seul graphique
plt.figure(figsize=(15, 10))
sns.boxplot(data=df[numeric_columns])
plt.xticks(rotation=90)  # Rotation des étiquettes pour qu'elles soient lisibles
plt.title("Boxplots des variables numériques (quantitatives et catégorielles
encodées)")
plt.show()
```

```
#Resumés statistiques
import pandas as pd
# Supposons que 'encoded_df' soit votre DataFrame contenant les variables encodées
et quantitatives
# Liste des variables quantitatives
numeric_columns = [
    'id', 'age', 'pression_academique', 'pression_liee_au_travail',
    'moyenne_notes', 'satisfaction_etudes', ' satisfaction_travail',
    'nombre_heure_travail_etude', 'stress_financier'
]
# Fonction pour calculer les statistiques descriptives
def calculate_statistics(df, numeric_columns):
    stats = {}
    for col in numeric_columns:
        stats[col] = {
            'Moyenne': df[col].mean(),
            'Médiane': df[col].median(),
            'Écart-type': df[col].std(),
            'IQR': df[col].quantile(0.75) - df[col].quantile(0.25),
            'Minimum': df[col].min(),
            'Maximum': df[col].max()
        }
    return pd.DataFrame(stats)
# Calculez les statistiques descriptives
statistics_df = calculate_statistics(df, numeric_columns)
statistics_df

import pandas as pd

# Supposons que 'encoded_df' soit votre DataFrame contenant les variables encodées
et quantitatives
# Liste des variables quantitatives
numeric_columns = [
    'id', 'age', 'pression_academique', 'pression_liee_au_travail',
    'moyenne_notes', 'satisfaction_etudes', ' satisfaction_travail',
    'nombre_heure_travail_etude', 'stress_financier'
]
# Fonction pour calculer les statistiques descriptives
def calculate_statistics(df, numeric_columns):
    stats = {}
    for col in numeric_columns:
        stats[col] = {
            'Moyenne': df[col].mean(),
            'Médiane': df[col].median(),
            'Écart-type': df[col].std(),
            'IQR': df[col].quantile(0.75) - df[col].quantile(0.25),
            'Minimum': df[col].min(),
            'Maximum': df[col].max()
        }
    return pd.DataFrame(stats)
# Calculez les statistiques descriptives
statistics_df = calculate_statistics(df, numeric_columns)
statistics_df

import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm
from statsmodels.stats.proportion import proportion_confint
```

```
# Création du DataFrame pour le tableau de contingence
data = {
    'depression': [0, 0, 1, 1],
    'pensees_suicidaire': ['No', 'Yes', 'No', 'Yes'],
    'count': [7866, 3699, 2379, 13957]
}
df_contingence = pd.DataFrame(data)

# Calcul de la proportion et de l'intervalle de confiance
total_students = 27901
total_suicidal = 17656
prop_suicidal = total_suicidal / total_students
conf_int = proportion_confint(total_suicidal, total_students, alpha=0.05)

# Visualisation de la courbe de l'intervalle de confiance
mean = prop_suicidal
std_err = np.sqrt((prop_suicidal * (1 - prop_suicidal)) / total_students)
x = np.linspace(mean - 4*std_err, mean + 4*std_err, 1000)
y = norm.pdf(x, mean, std_err)

plt.figure(figsize=(10, 6))
plt.plot(x, y, label='Distribution de la proportion')
plt.fill_between(x, y, where=((x >= conf_int[0]) & (x <= conf_int[1])),
color='skyblue', alpha=0.4, label='Intervalle de confiance à 95%')
plt.axvline(x=mean, color='blue', linestyle='--', label='Proportion de pensées
suicidaires')
plt.xlabel('Proportion')
plt.ylabel('Densité')
plt.title('Intervalle de confiance pour la proportion d\'étudiants avec pensées
suicidaires')
plt.legend()
plt.show()
```