MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE RECHERCHE SCIENTIFIQUE

REPUBLIQUE DE COTE D'IVOIRE

Institut Supérieur de Statistique D'Econométrie et de Data Science

Union-Discipline-Travail

MASTER 1

STATISTIQUES – ECONOMETRIE – DATA SCIENCE

MINI PROJET

ANALYSE STATISTIQUES INFERENTIELLE

ANALYSE DES FACTEURS INFLUENÇANT LA DÉPRESSION ET LA SANTÉ MENTALE DES ÉTUDIANTS

ANNEE ACADEMIQUE

2024 – 2025

Nom: YOBO

Prenom(s): BAYE GUY ANGE HENOC

Enseignant – Encadreur

AKPOSSO DIDIER MARTIAL

AVANT-PROPOS

Ce document présente un ensemble de données sur la dépression chez les étudiants, visant à identifier les facteurs qui influencent leur bien-être mental. L'analyse prend en compte des éléments tels que l'âge, le sexe, les performances académiques, les habitudes de vie et les antécédents de santé mentale. Des échelles de dépression standardisées permettent également d'évaluer les niveaux de dépression parmi les étudiants.

Ces données sont utiles pour les chercheurs, psychologues et professionnels de l'éducation afin d'identifier les facteurs de risque et d'élaborer des stratégies de prévention et de soutien. L'objectif est de mieux comprendre les causes de la dépression chez les étudiants et de proposer des solutions adaptées pour améliorer leur santé mentale.

Table des matières

Liste des tableaux

Liste des figures

INTRODUCTION

La santé mentale des étudiants est un enjeu de plus en plus important dans le monde académique et social. La dépression, en particulier, constitue l'un des troubles les plus répandus parmi cette population, affectant non seulement leur bien-être psychologique, mais aussi leurs performances académiques et leur vie sociale. Cette étude se concentre sur l'analyse des facteurs influençant la dépression chez les étudiants, en explorant les tendances et les éléments explicatifs pouvant expliquer les variations de ces niveaux de dépression.

La problématique principale de cette étude réside dans : ***le besoin de mieux comprendre les facteurs contributifs à la dépression chez les étudiants, afin de pouvoir identifier des stratégies efficaces pour prévenir ce phénomène et améliorer la santé mentale au sein de cette population.*** En effet, des éléments tels que les caractéristiques démographiques, les habitudes de vie, les performances académiques et les antécédents familiaux de santé mentale peuvent avoir un impact significatif sur les risques de développer une dépression.

Les principaux résultats attendus de cette analyse sont d'identifier les facteurs les plus influents dans l'apparition de la dépression chez les étudiants, ainsi que de mieux comprendre la relation entre ces facteurs. Cette étude vise à fournir des recommandations pour des interventions ciblées, notamment en matière de soutien psychologique, de gestion du stress et d'amélioration du bien-être des étudiants.

La méthodologie adoptée repose sur une série d'analyses statistiques appliquées à un ensemble de données comprenant des informations démographiques, académiques et de santé. Les techniques de prétraitement incluent le nettoyage et la normalisation des données, afin de préparer les données pour les analyses ultérieures. Les principales analyses statistiques comprennent des Analyse de la Corrélation, Régression Linéaire Simple et Multiple, Test d'Hypothèses, Analyse de Variance (ANOVA), Test Chi-Carré.

Ces analyses reposent sur des fondements théoriques de la psychologie et des sciences sociales, en particulier les modèles de stress et de santé mentale, pour en tirer des conclusions significatives et pertinentes.

Ainsi, cette étude vise à éclairer la compréhension des facteurs influençant la dépression chez les étudiants, en utilisant des approches statistiques rigoureuses, et à offrir des pistes concrètes pour des interventions préventives efficaces.

I) APPROCHE METHODOLOGIQUE DES DONNEES

L'**approche méthodologique de données** consiste à l'organisation, la collecte, l'analyse et l'interprétation des données dans le cadre d'une étude ou d'une recherche. Cela implique l'ensemble des principes, techniques et processus qui permettent de traiter les données de manière systématique et rigoureuse pour parvenir à des résultats fiables et pertinents.

1) Présentation du dictionnaire de données

Un dictionnaire de données est un document qui fournit une description exhaustive de chaque variable utilisée dans une analyse statistique ou économétrique. Il précise les propriétés, les caractéristiques et le contexte de chaque variable, tout en clarifiant leur signification et leur rôle dans l'analyse.

id	Identifiant	Identifiant unique de chaque étudiant.	Identifiant numérique unique (ex. : 1, 2, 3, ...)
sexe	Qualitative (nominale)	Sexe de l'étudiant.	Masculin, Féminin
age	Quantitative (continue)	Âge de l'étudiant.	Valeurs numériques (ex. : 18, 20, 25, ...)
ville	Qualitative (nominale)	Ville de résidence de l'étudiant.	Liste de villes possibles (ex. : Paris, Lyon, Marseille, ...)
profession	Qualitative (nominale)	Profession de l'étudiant (étudiant à temps plein ou autre activité).	Étudiant à temps plein, Autre activité
pression_academique	Quantitative (discrète)	Niveau de pression académique ressenti par l'étudiant (échelle).	Valeurs sur une échelle (ex. : 1-10, où 1 = faible pression et 10 = forte pression)
pression_liee_au_travail	Quantitative (discrète)	Niveau de pression liée au travail ressenti par l'étudiant (échelle).	Valeurs sur une échelle (ex. : 1-10, où 1 = faible pression et 10 = forte pression)
moyenne_notes	Quantitative (continue)	Moyenne générale des notes obtenues par l'étudiant.	Valeurs numériques (ex. : 10.5, 15.8, ...)
satisfaction_etudes	Quantitative (discrète)	Niveau de satisfaction de l'étudiant par rapport à ses études.	Valeurs sur une échelle (ex. : 1-10, où 1 = très insatisfait et 10 = très satisfait)
satisfaction_travail	Quantitative (discrète)	Niveau de satisfaction de l'étudiant par rapport à son travail.	Valeurs sur une échelle (ex. : 1-10, où 1 = très insatisfait et 10 = très satisfait)
duree_sommeil	Qualitative (ordinale)	Durée du sommeil de l'étudiant, catégorielle.	Moins de 5 heures, 5-6 heures, 7-8 heures, plus de 8 heures
habitudes_alimentaires	Qualitative (nominale)	Type d'habitudes alimentaires de l'étudiant.	Saines, Modérées, Mauvaises
diplome_suivi	Qualitative (nominale)	Diplôme suivi ou obtenu par l'étudiant.	BSc, M.Tech, MSc, PhD, Autre
pensees_suicidaires	Qualitative (nominale)	Si l'étudiant a déjà eu des pensées suicidaires.	Oui, Non
nombre_heure_travail_etude	Quantitative (discrète)	Nombre d'heures de travail ou d'études par jour de l'étudiant.	Valeurs numériques (ex. : 3, 5, 8, ...)
stress_financier	Quantitative (discrète)	Niveau de stress financier ressenti par l'étudiant (échelle ou score).	Valeurs sur une échelle (ex. : 1-10, où 1 = faible stress et 10 = fort stress)
antecedents_familiaux_maladies_mentales	Qualitative (nominale)	Présence ou absence d'antécédents familiaux de maladies mentales.	Oui, Non
depression	Qualitative (nominale)	Si l'étudiant souffre de dépression (1 pour oui, 0 pour non).	1 (oui), 0 (non)

Tableau 1Dictionnaire des données

2) Présentation du jeu de donnée

id	sexe	age	ville	profession	pression_academique	pression_liee_au_travail	moyenne_notes	satisfaction_etudes	satisfaction_travail	duree_sommeil	habitudes_alimentaires	diplome_suivi	pensees_suicidaire	nombre_heure_travail_etu	stress_financier	antecedents_familiaux_mal	depression
0	Male	33.0	Visakhapatna	Student	5.0	0.0	8.97	2.0	0.0	5-6 hours	Healthy	B.Pharm	Yes	3.0	1.0	No	1
1	Femal	24.0	Bangal	Studen	2.0	0.0	5.90	5.0	0.0	5-6	Moder	BSc	No	3.0	2.0	Yes	0
2	Male	31.0	Srinagar	Student	3.0	0.0	7.03	5.0	0.0	Less than 5	Healthy	BA	No	9.0	1.0	Yes	0
3	Femal	28.0	Varana	Studen	3.0	0.0	5.59	2.0	0.0	7-8	Moder	BCA	Yes	4.0	5.0	Yes	1
4	Femal	25.0	Jaipur	Studen	4.0	0.0	8.13	3.0	0.0	5-6	Moder	M.Tech	Yes	1.0	1.0	No	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
27896	Femal	27.0	Surat	Studen	5.0	0.0	5.75	5.0	0.0	5-6	Unheal	Class	Yes	7.0	1.0	Yes	0
27897	Male	27.0	Ludhiana	Student	2.0	0.0	9.40	3.0	0.0	Less than 5	Healthy	MSc	No	0.0	3.0	Yes	0
27898	Male	31.0	Farida	Studen	3.0	0.0	6.61	4.0	0.0	5-6	Unheal	MD	No	12.0	2.0	No	0
27899	Female	18.0	Ludhiana	Student	5.0	0.0	6.88	2.0	0.0	Less than 5	Healthy	Class 12	Yes	10.0	5.0	No	1
27900	Male	27.0	Patna	Student	4.0	0.0	9.24	1.0	0.0	Less than 5	Healthy	BCA	Yes	2.0	3.0	Yes	1

Tableau 2 Extrait du jeu de donnée

L'ensemble de données présente 27901 observations et 18 variables. Il semble qu'il y ait des valeurs manquantes dans les données qui n'influence pas le déroulé de l'analyse.

2.1) ***Présentation d'information supplémentaire sur le jeu de données***

Variable	Nature	Valeur
habitudes_alimentaires	Unhealthy	10317
	Moderate	9921
	Healthy	7651
	Others	12
diplome_suivi	Class 12	6080
	B.Ed	1867
	B.Com	1506
	B.Arch	1478
	BCA	1433
	MSc	1190
	B.Tech	1152
	MCA	1044
	M.Tech	1022
	BHM	925
pensees_suicidaire	Yes	17656
	No	10245
antecedants_familiaux_maladie_mentale	No	14398
	Yes	13503

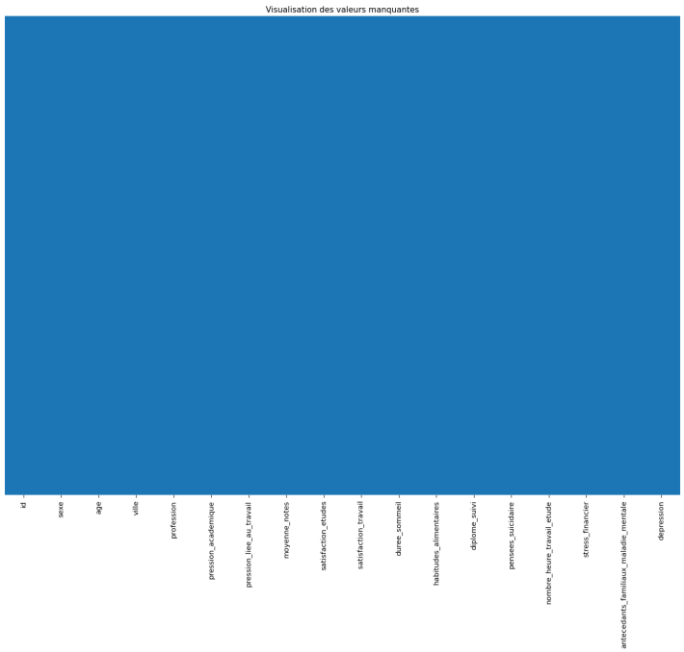
Variable	Nature	Valeur
sexe	Male	15547
	Female	12354
ville	Kalyan	1570
	Srinagar	1372
	Hyderabad	1340
	Vasai-Virar	1290
	Lucknow	1155
	Thane	1139
	Ludhiana	1111
	Agra	1094
	Surat	1078
	Kolkata	1066
profession	Student	27870
	Architect	8
	Teacher	6
	Digital Marketer	3
	Content Writer	2
	Chef	2
	Doctor	2
	Pharmacist	2
	Civil Engineer	1
	UX/UI Designer	1
duree_sommeil	Less than 5 hours	8310
	7-8 hours	7346
	5-6 hours	6183
	More than 8 hours	6044
	Others	18

Tableau 3 Count des valeurs présente

3) Détection et apurement des valeurs manquantes

Dans cette partie, nous allons essayer de détecter visuellement si toute fois notre jeu de donnée présente des valeurs manquantes, et les traiter.

Ces valeurs peuvent résulter d'erreurs de mesure, de saisie, de calcul ou simplement de valeurs extrêmes réelles dans les données. Les valeurs atypiques peuvent avoir un impact considérable sur les analyses statistiques en influençant les résultats des statistiques descriptives comme la moyenne et l'écart-type, ainsi que sur les tests d'hypothèse. Il est donc essentiel de repérer et de gérer ces valeurs extrêmes avant d'entreprendre une analyse statistique.



La figure ci-contre présente l'ensemble des valeurs manquantes, elle n'est pratiquement pas visible du fait de sa faible quantité.

Ces valeurs demeureront pour la suite de nos analyses

Figure 1 Visualisation des valeurs manquantes

II) ANALYSE EXPLORATOIRE DES DONNEES

L'analyse exploratoire des données (ou **EDA** pour *Exploratory Data Analysis*) constitue une étape essentielle dans le processus d'analyse de données. Son objectif principal est d'examiner et de comprendre les caractéristiques d'un jeu de données avant l'application de modèles statistiques ou d'algorithmes complexes. Cette phase permet non seulement de formuler des hypothèses, mais aussi d'identifier des patterns, des anomalies et des relations potentielles entre les variables. Elle joue également un rôle clé dans la préparation des données pour des analyses plus approfondies.

1) Analyse univariée

Dans cette partie nous aborderons une analyse systématique sur chaque variable de surcroît la variable d'intérêt.

1.1) Paramètre statistiques (SKEWNESS)

Variable	Skewness	Interprétation
age	0.132239	L'asymétrie est faible et positive, ce qui signifie que la distribution de l'âge est légèrement inclinée vers la droite, mais elle est proche de la symétrie.
pression_academique	-0.135165	L'asymétrie est faible et négative, ce qui suggère une légère inclinaison vers la gauche, mais cela reste relativement équilibré.
pression_liee_au_travail	108.594361	L'asymétrie est très élevée et positive, ce qui indique une forte concentration des valeurs faibles et une longue queue vers les valeurs élevées. Cela suggère que la plupart des individus ressentent une pression académique relativement faible, mais quelques-uns en ressentent une pression extrême.
moyenne_notes	-0.113063	L'asymétrie est faible et négative, ce qui suggère une légère inclinaison vers la gauche. Cela peut signifier que la majorité des étudiants ont des moyennes relativement élevées, mais il existe quelques notes très faibles.
satisfaction_etudes	0.010423	L'asymétrie est très faible et positive, indiquant que la distribution de la satisfaction des études est pratiquement symétrique.
satisfaction_travail	74.105663	Comme pour la pression liée au travail, l'asymétrie est extrêmement élevée et positive, suggérant que la majorité des individus ont une faible satisfaction au travail, tandis que quelques-uns en sont extrêmement satisfaits.
nombre_heure_travail_etude	-0.454769	L'asymétrie est modérément négative, indiquant une inclinaison vers la gauche. Cela pourrait signifier que la plupart des étudiants travaillent ou étudient pendant un nombre modéré d'heures, mais il existe des étudiants qui travaillent ou étudient très peu.
stress_financier	-0.130304	L'asymétrie est faible et négative, indiquant une légère inclinaison vers la gauche, ce qui peut signifier que la plupart des individus ne ressentent qu'un stress financier modéré.

Tableau 4 Paramètre statistiques

1.2) Représentation graphique des variables

a) Distribution des variable numériques

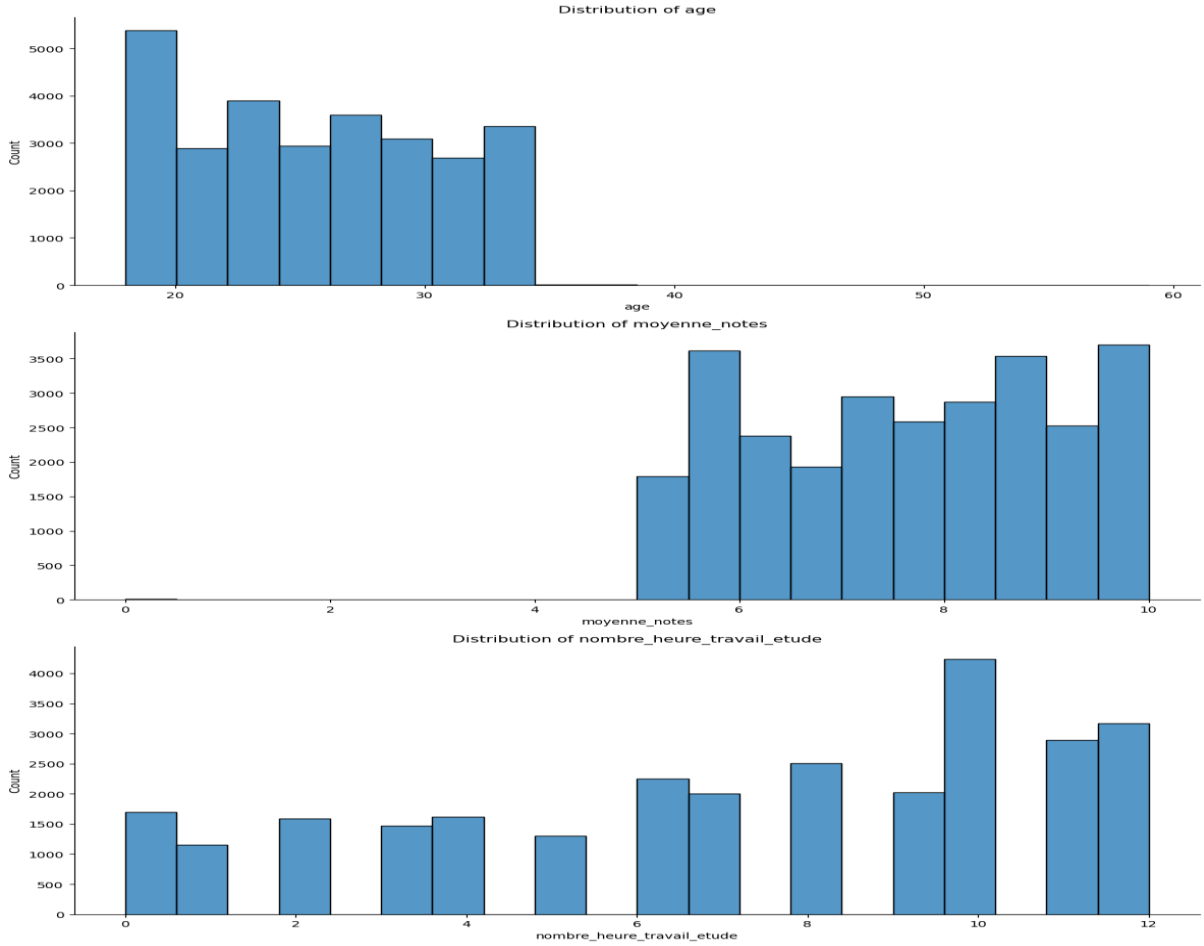


Figure 2 Variables numériques

D'après l'analyse de l'image, on peut observer trois graphiques à barres montrant la distribution de différentes variables.

Le premier graphique représente la distribution de l'âge, avec une concentration plus élevée autour de 20 ans et une diminution du nombre d'individus à mesure que l'âge augmente.

Le deuxième graphique montre la distribution de la moyenne des notes, avec un pic autour d'une moyenne de 4 et une diminution du nombre d'individus aux extrémités.

Enfin, le troisième graphique illustre la distribution du nombre d'heures de travail d'étude, avec une concentration plus élevée autour de 10 heures et une diminution du nombre d'individus avec des nombres d'heures plus faibles ou plus élevés. Ces distributions peuvent fournir des informations importantes sur la répartition des données et les tendances observées dans chaque variable.

b) Distribution des caractéristiques catégorielles

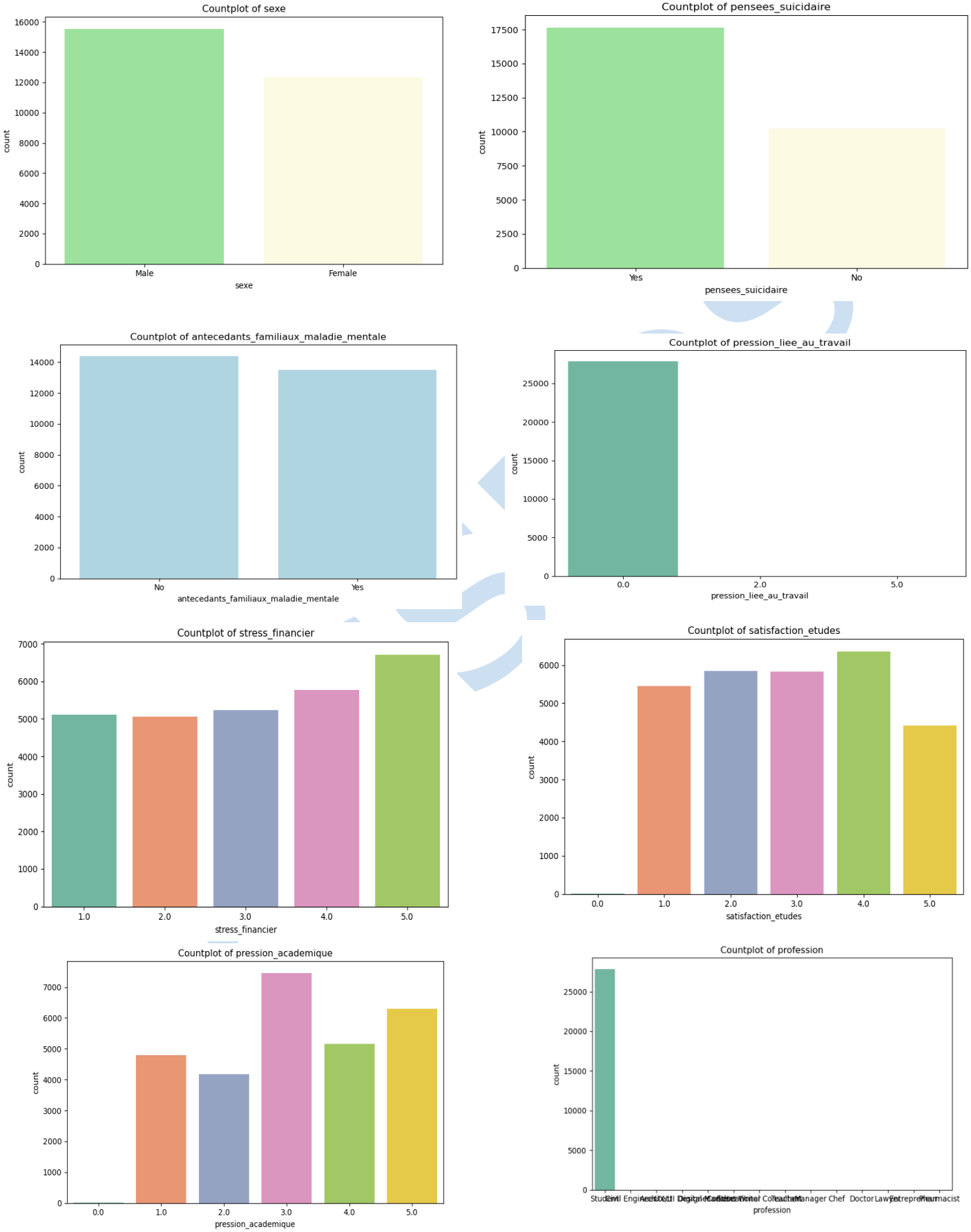


Figure 3Variables catégorielles

Le graphique **"Countplot of antecedents_familiaux_maladie_mentale"** montre une distribution équilibrée entre les réponses "Yes" et "No" en ce qui concerne les antécédents familiaux de maladie mentale. Cela suggère que le nombre de personnes ayant des antécédents familiaux de maladie mentale est à peu près égal à celles n'en ayant pas.

Le graphique **"Countplot of sexes"** indique qu'il y a un nombre plus élevé de personnes de sexe masculin que de personnes de sexe féminin. Cela peut être utile pour comprendre la répartition des sexes dans l'échantillon étudié.

Le graphique **"Countplot of pensees_suicidaire"** montre un nombre plus élevé de personnes avec la réponse "No" par rapport à "Yes" en ce qui concerne les pensées suicidaires. Cela suggère que la majorité des personnes dans l'échantillon n'ont pas de pensées suicidaires.

Le graphique **"Countplot of pression_liée_au_travail"** révèle un nombre plus élevé de personnes signalant une pression liée au travail. Cela peut indiquer un niveau élevé de stress ou de pression dans l'environnement de travail des individus.

Le graphique **"Countplot of satisfaction_etudes"** montre une distribution équilibrée entre les différents niveaux de satisfaction en matière d'études. Cela suggère que les individus de l'échantillon ont des niveaux variés de satisfaction par rapport à leurs études.

Le graphique **"Countplot of pression_academique"** montre une distribution équilibrée entre les différents niveaux de pression académique. Cela indique que les individus de l'échantillon ressentent des niveaux variés de pression liée à leurs études.

Le graphique **"Countplot of profession"** montre une distribution équilibrée entre les différentes professions. Cela peut être utile pour comprendre la diversité des professions représentées dans l'échantillon.

2) Analyse bivariée

Dans le cadre de l'analyse bivariées nous tenterons de voir s'il y a une relation entre les variable catégorielles et numériques.

2.1) ***Distribution des caractéristiques numériques selon la présence de dépression***

La "distribution des caractéristiques numériques selon la présence de dépression" fait référence à une analyse visant à examiner la répartition de diverses caractéristiques quantitatives en fonction de la présence ou de l'absence de dépression au sein d'un groupe de personnes.

Les caractéristiques numériques, telles que l'âge, le niveau de stress, la qualité du sommeil, ou encore les résultats de certains tests psychométriques ou physiologiques, sont analysées pour observer leur variabilité et leur comportement dans les deux sous-groupes : ceux présentant des symptômes de dépression et ceux n'en présentant pas.

a) Représentations graphiques

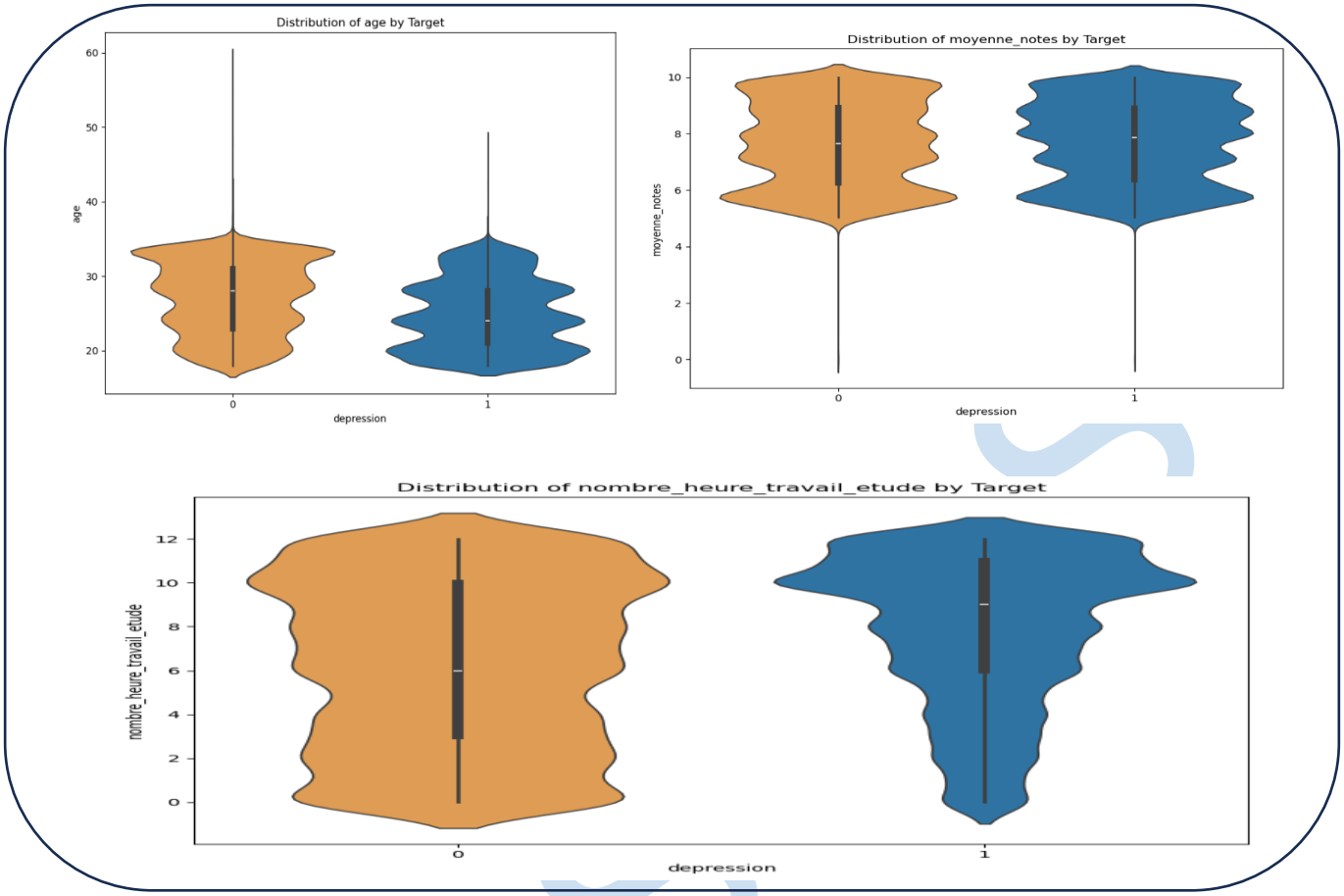


Figure 4 Distribution numériques selon la dépression

Le premier graphique montre la distribution de l'âge pour les deux cibles. On peut observer que la distribution de l'âge pour les individus souffrant de dépression et ceux sans dépression est différente. Par exemple, il semble y avoir une concentration plus élevée de jeunes dans le groupe de dépression par rapport au groupe sans dépression.

Le deuxième graphique représente la distribution de la moyenne des notes en fonction de la cible. On peut voir que la distribution des notes moyennes pour les individus souffrant de dépression et ceux sans dépression est également différente. Il semble y avoir une tendance à des notes plus basses pour le groupe de dépression par rapport au groupe sans dépression.

Le troisième graphique montre la distribution du nombre d'heures de travail d'étude en fonction de la cible. On peut observer des différences dans la distribution du nombre d'heures passées au travail et à l'étude entre les deux groupes. Par exemple, il pourrait y avoir une concentration plus élevée d'heures de travail et d'étude pour le groupe de dépression par rapport au groupe sans dépression.

2.2) ***Relation entre les colonnes catégorielles et une variable cible (Dépression)***

Figure 5 Distributions catégorielles selon la dépression

Graphique 1: Stacked Bar Plot of sexe and Target

Ce graphique montre la répartition de la dépression en fonction du sexe. On peut voir que les hommes ont un taux de dépression plus élevé que les femmes.

Graphique 2: Stacked Bar Plot of pensees suicidaire and Target

Ce graphique illustre les pensées suicidaires en fonction de la cible. On remarque que les personnes ayant des pensées suicidaires sont plus nombreuses que celles n'en ayant pas.

Graphique 3: Stacked Bar Plot of antecedants familiaux maladie mentale and Target

Ce graphique présente les antécédents familiaux de maladie mentale en fonction de la cible. On peut observer que les personnes ayant des antécédents familiaux de maladie mentale sont plus nombreuses que celles n'en ayant pas.

Graphique 4: Stacked Bar Plot of pression academique and Target

Ce graphique montre la pression académique en fonction de la cible. On remarque que les personnes subissant une forte pression académique sont plus nombreuses que celles n'en subissant pas.

Graphique 5: Stacked Bar Plot of satisfaction etudes and Target

Ce graphique illustre la satisfaction des études en fonction de la cible. On peut voir que les personnes satisfaites de leurs études sont plus nombreuses que celles qui ne le sont pas.

Graphique 6: Stacked Bar Plot of stress financier and Target

Ce graphique présente le stress financier en fonction de la cible. On remarque que les personnes subissant un fort stress financier sont plus nombreuses que celles n'en subissant pas.

Graphique 7: Stacked Bar Plot of profession and Target

Ce graphique montre la profession en fonction de la cible. On peut observer que les professions les plus touchées par la dépression sont les étudiants, les enseignants et les entrepreneurs.

Graphique 8: Stacked Bar Plot of pression liee au travail and Target

Ce graphique illustre la pression liée au travail en fonction de la cible. On remarque que les personnes subissant une forte pression liée au travail sont plus nombreuses que celles n'en subissant pas.

2.3) Création du treemap des professions

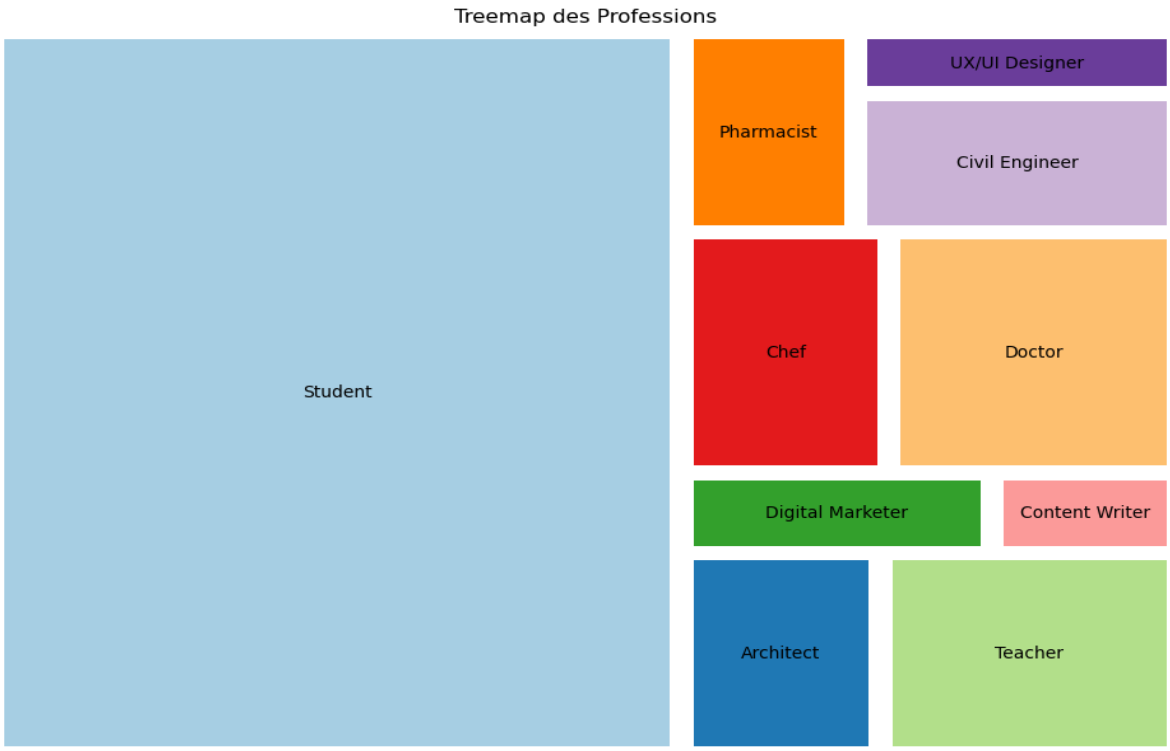


Figure 6 Treemap des professions

2.4) Les principales professions et leur relation avec la dépression

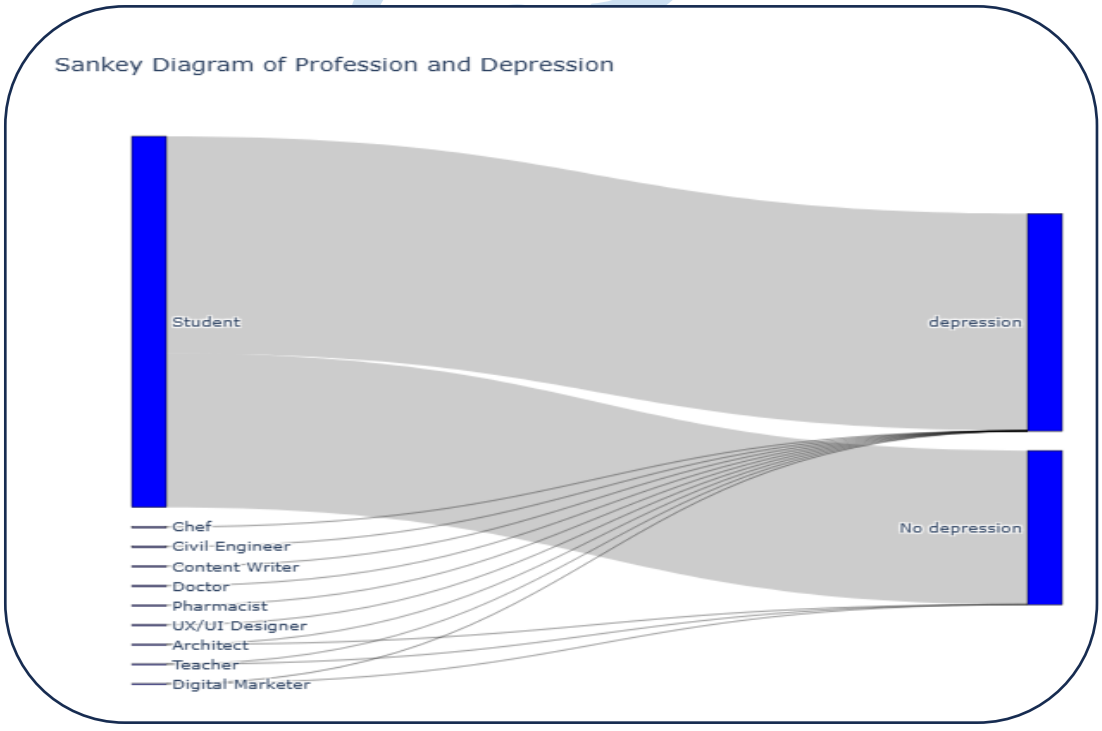


Figure 7 Relation profession et dépression

Ce diagramme illustre de manière visuelle la répartition de la dépression en fonction des différentes professions. Chaque profession est représentée par une ligne, dont la largeur est proportionnelle au nombre de personnes dans cette profession.

On peut observer que :

- Les professions les plus touchées par la dépression sont les étudiants, les enseignants et les entrepreneurs. Leurs lignes sont plus larges dans la partie "dépression" du diagramme.
- Les professions les moins touchées par la dépression semblent être les chefs, les ingénieurs civils, les rédacteurs de contenu, les médecins et les pharmaciens. Leurs lignes sont plus fines dans la partie "dépression".
- Certaines professions comme les architectes, les concepteurs UX/UI et les spécialistes du marketing digital présentent un mélange de personnes avec et sans dépression.

2.5) Création du treemap des diplômes suivis

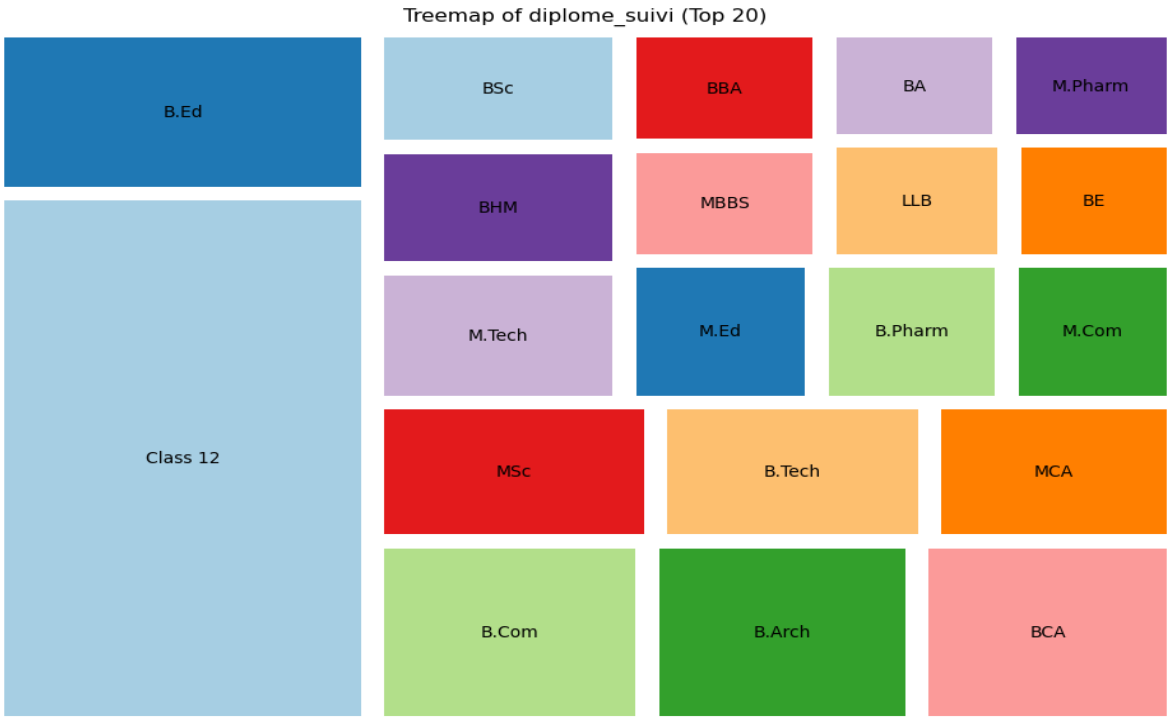


Figure 8 Treemap des diplômes

2.6) Les meilleurs diplôme suivi et leur relation avec la dépression

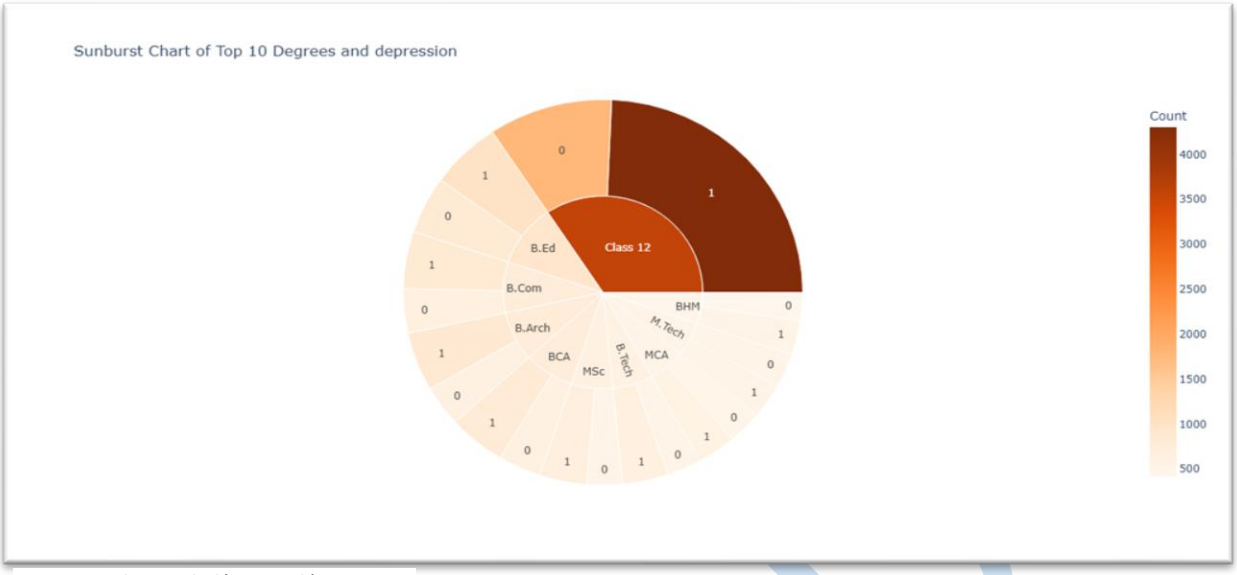


Figure 9 Relation diplôme et dépression

- Le diplôme le plus répandu est le B.Ed (Bachelor of Education), avec 1 personne sans dépression et 1 personne avec dépression.
- Le diplôme de Class 12 (équivalent du baccalauréat) a 1 personne avec dépression.
- Le diplôme de B.Com (Bachelor of Commerce) a 1 personne sans dépression.
- Les diplômes de BHM (Bachelor of Hotel Management) et B.Arch (Bachelor of Architecture) ont chacun 1 personne sans dépression.
- Le diplôme de M.Tech (Master of Technology) a 1 personne avec dépression.
- Les diplômes de BCA (Bachelor of Computer Applications), MCA (Master of Computer Applications) et MSc (Master of Science) ont chacun 1 personne sans dépression.
- Le diplôme de B.Tech (Bachelor of Technology) n'a pas de personne représentée, que ce soit avec ou sans dépression.

2.7) ***Distribution sur la variable dépression***

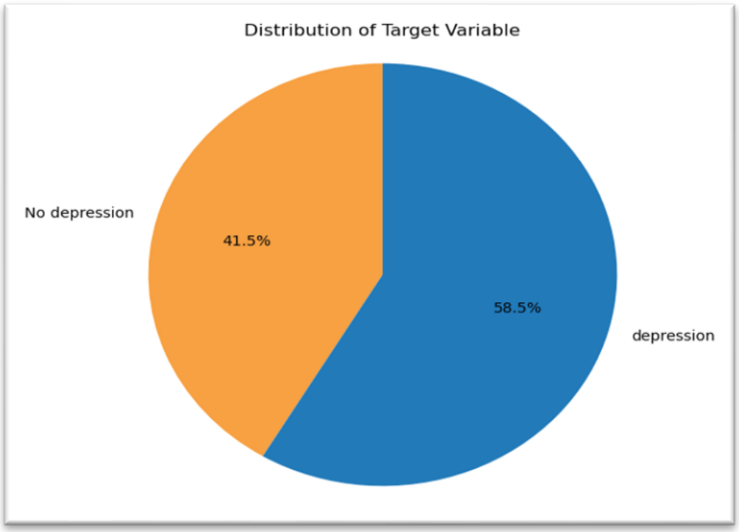


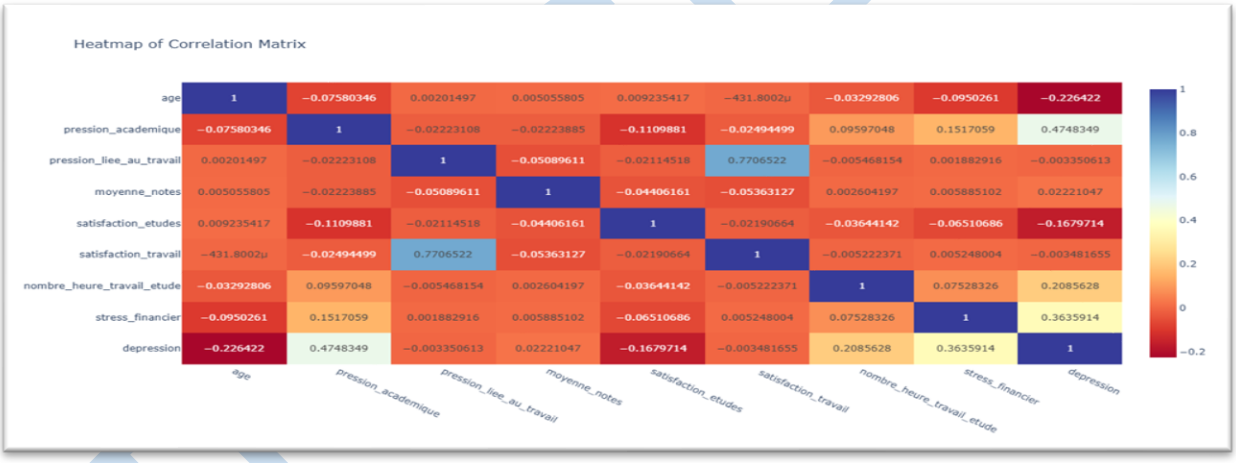
Figure 10 Visualisation de la dépression

D'après le graphique en camembert, la distribution de la variable cible est la suivante :

- 58,5% des observations sont dans la catégorie "dépression"
- 41,5% des observations sont dans la catégorie "No dépression"

Cela indique que la majorité des observations dans ce jeu de données sont dans la catégorie "dépression".

2.8) ***Corrélation entre les variables***



Cette carte thermique représente la matrice de corrélation entre les différentes variables du jeu de données. Voici les principales observations :

- La variable "age" a une corrélation négative avec la plupart des autres variables, en particulier avec "satisfaction_travail".
- La variable "pression_academique" a une forte corrélation positive avec "depression" et une corrélation négative avec "satisfaction_etudes".
- La variable "pression_liee_au_travail" a une forte corrélation positive avec "satisfaction_travail".
- La variable "moyenne_notes" a une corrélation positive avec "satisfaction_etudes" et "stress_financier".
- La variable "nombre_heure_travail_etude" a une corrélation positive avec "stress_financier" et "depression".

Figure 11 visualisation des deux matrices de correlation

Analyse détaillée de la matrice de corrélation

La carte thermique montre les coefficients de corrélation entre les différentes variables du jeu de données. Voici les principales observations :

Corrélations positives

- La variable "pression_liée_au_travail" a une forte corrélation positive (>0,7) avec la variable "satisfaction_travail". Cela indique que plus la pression liée au travail est élevée, plus la satisfaction au travail est importante.
- La variable "nombre_heure_travail_etude" a une corrélation positive (>0,05) avec la variable "stress_financier". Cela suggère que plus les heures de travail et d'études sont importantes, plus le stress financier est élevé.
- La variable "depression" a une corrélation positive (>0,35) avec la variable "stress_financier". Cela montre que le stress financier est lié à la dépression.

Corrélations négatives

- La variable "age" a des corrélations négatives avec la plupart des autres variables, notamment avec "satisfaction_travail" (-0,43). Cela indique que plus l'âge augmente, moins la satisfaction au travail est élevée.
- La variable "pression_académique" a une forte corrélation négative (-0,11) avec "satisfaction_etudes". Cela suggère que plus la pression académique est importante, moins la satisfaction dans les études est élevée.
- La variable "moyenne_notes" a une corrélation négative (-0,05) avec "satisfaction_etudes". Cela montre que de meilleures notes ne sont pas nécessairement liées à une plus grande satisfaction dans les études.

ANALYSE STATISTIQUE INFERENTIELLE

L'analyse inférentielle est une branche des statistiques qui permet de faire des inférences ou des généralisations à partir de données observées. Contrairement à l'analyse descriptive, qui se concentre sur la présentation et la description des caractéristiques d'un ensemble de données, l'analyse inférentielle utilise des échantillons pour faire des déductions sur une population entière.

Les principales techniques d'analyse inférentielle incluent :

1. **Estimation des paramètres** : Il s'agit d'estimer des paramètres inconnus de la population, comme la moyenne ou l'écart-type, à partir d'un échantillon. Cette estimation peut être ponctuelle (par exemple, la moyenne échantillonnale) ou par intervalle (comme un intervalle de confiance autour de la moyenne).
2. **Tests d'hypothèses** : Ces tests permettent de vérifier si une hypothèse sur une population est valide, en utilisant des échantillons. Par exemple, on peut tester si la moyenne d'un échantillon est égale à une valeur hypothétique (test de la moyenne), ou tester l'égalité de deux moyennes dans deux groupes différents (test de Student).
3. **Analyse de la variance (ANOVA)** : Cette méthode permet de tester si les moyennes de plusieurs groupes sont significativement différentes les unes des autres.
4. **Régression et corrélation** : Ces techniques permettent d'examiner les relations entre différentes variables. La régression linéaire, par exemple, cherche à prédire la valeur d'une variable en fonction d'une ou plusieurs autres variables.
5. **Estimation bayésienne** : Cette approche repose sur le théorème de Bayes et permet d'inférer des probabilités conditionnelles à partir de données observées.

Processus de l'analyse inférentielle

1. **Formulation de l'hypothèse** : Une hypothèse nulle (H0) et une hypothèse alternative (H1) sont formulées pour tester une assertion sur la population.
2. **Choix du test statistique** : Selon la nature des données et de l'hypothèse, un test statistique (t-test, chi carré, ANOVA, etc.) est choisi.
3. **Calcul de la statistique de test** : À partir des données de l'échantillon, on calcule une statistique qui permet de déterminer si l'hypothèse nulle doit être rejetée ou non.
4. **Prise de décision** : En fonction du résultat du test (p-valeur), on décide si l'hypothèse nulle peut être rejetée, ce qui implique que l'effet observé dans l'échantillon est significatif.

III) ESTIMATIONS STATISTIQUES

1) **Intervalle de confiance :** proportion d'étudiants ayant eu des pensées suicidaires.

L'intervalle de confiance donne une plage de valeurs dans laquelle on s'attend à ce que le vrai paramètre de la population se trouve, basé sur un échantillon observé.

Intervalle de confiance
IC = (0.6272, 0.6385)
Soit : 62.72% et 63.85%.

Tableau 5 intervalle de confiance pensée suicidaire

Cela signifie que, avec un niveau de confiance de 95%, nous pouvons affirmer que la proportion réelle d'étudiants dans la population ayant eu des pensées suicidaires se situe quelque part entre **62.72% et 63.85%**.

2) Moyenne et Médiane :

2.1) *Heures de travail ou d'études pour les étudiants souffrant de dépression*

Moyenne Heures de travail ou d'études pour les étudiants	Médiane Heures de travail ou d'études pour les étudiants
7.81 heures	**9.00 heures**

Tableau 6 Moyenne et médiane des heures de travails ou d'études

Moyenne des heures de travail/études pour les étudiants dépressifs : 7.81 heures

- La **moyenne** représente la valeur centrale des heures de travail/études pour tous les étudiants dépressifs. En d'autres termes, la moyenne de **7.81 heures** signifie que, en moyenne, un étudiant dépressif consacre environ 7.81 heures par semaine à son travail/études.

Médiane des heures de travail/études pour les étudiants dépressifs : 9.00 heures

- La **médiane** est la valeur qui sépare les données en deux parties égales. Cela signifie que **50%** des étudiants dépressifs passent moins de **9 heures** à travailler/étudier, et l'autre **50%** passent plus de **9 heures**.

Conclusion :

- La **moyenne (7.81 heures)** étant inférieure à la **médiane (9.00 heures)**, cela indique que la distribution des heures de travail/études est **asymétrique**, avec quelques étudiants travaillant peu, ce qui abaisse la moyenne. Une **médiane plus élevée que la moyenne** suggère une **asymétrie positive**, où une majorité d'étudiants dépressifs consacre plus de temps aux études.

2.2) ***Stress financier : pour les étudiants avec et sans dépression***

L'objectif global de cette étude est de mieux comprendre les liens entre la situation financière et la santé mentale des étudiants, afin de proposer des solutions adaptées pour améliorer leur bien-être et leur réussite académique.

a) Les étudiants avec dépression

mean_stress_depressed	median_stress_depressed
3.58	**4.00**

Tableau 7moyenne et médiane de stresse financier (dépression)

1. Moyenne du stress financier pour les étudiants dépressifs : 3.58

- La **moyenne** de **3.58** indique que, en moyenne, les étudiants dépressifs ressentent un certain niveau de stress financier, sur une échelle donnée (probablement de 1 à 5 ou 1 à 7, en fonction des données). Cela suggère que le stress financier n'est pas extrême, mais assez notable chez ces étudiants.

2. Médiane du stress financier pour les étudiants dépressifs : 4.00

- La **médiane** de **4.00** signifie que **50%** des étudiants dépressifs ressentent un stress financier inférieur à 4 et l'autre **50%** ressentent un stress supérieur à 4.
- La médiane donne une idée plus robuste de la tendance centrale, indépendamment des valeurs extrêmes.

Conclusion :

- La **moyenne** (3.58) est légèrement inférieure à la **médiane** (4.00), ce qui suggère que la distribution du stress financier pourrait être **asymétrique**, avec quelques étudiants ressentant un stress plus faible, ce qui abaisse la moyenne.
- Globalement, les étudiants dépressifs semblent éprouver un stress financier modéré, avec une tendance vers des niveaux de stress plus élevés pour la moitié d'entre eux.

b) Les étudiants sans dépression

mean_stress_not_depressed	median_stress_not_depressed
2.52	**2.00**

Tableau 8 Moyenne et médiane stresse financier (sans dépression)

Moyenne du stress financier pour les étudiants non dépressifs : 2.52

- La **moyenne** de **2.52** suggère que, en moyenne, les étudiants non dépressifs ressentent un **stress financier modéré**. Ce niveau est inférieur à celui des étudiants dépressifs, ce qui indique que le stress financier est généralement moins élevé chez les étudiants non dépressifs.

2. Médiane du stress financier pour les étudiants non dépressifs : 2.00

- La **médiane** de **2.00** signifie que **50%** des étudiants non dépressifs ressentent un stress financier inférieur à 2 et l'autre **50%** ressentent un stress supérieur à 2.
- La médiane plus basse que la moyenne suggère que la distribution du stress financier est **asymétrique**, avec une majorité d'étudiants non dépressifs ayant un stress financier relativement faible.

Conclusion :

- Les étudiants non dépressifs ont un **stress financier moyen de 2.52**, avec une **médiane de 2.00**. Cela indique que la plupart des étudiants non dépressifs ressentent un stress financier faible à modéré, tandis que quelques-uns peuvent ressentir un stress financier plus élevé, ce qui tire légèrement la moyenne vers le haut. En comparaison, le stress financier est plus élevé chez les étudiants dépressifs.

IV) TEST DE COMPARAISON DE POPULATIONS (TEST D'ÉGALITÉ DE MOYENNES)

L'objectif de ce test est de vérifier si la différence observée entre les moyennes de deux populations ou groupes (par exemple, étudiants dépressifs vs non dépressifs) est significative ou si elle pourrait être le résultat du hasard.

1) **Test d'égalité de moyennes** : *La satisfaction des études diffère-t-elle significativement entre les étudiants souffrant de dépression et ceux qui n'en souffrent pas ?*

Cette réflexion nous amène à examiner si la dépression peut influencer la perception des étudiants vis-à-vis de leur expérience académique, et si des différences notables existent entre les deux groupes, ce qui pourrait avoir des implications importantes sur les politiques de soutien et d'accompagnement des étudiants.

Tableau de contingence

Satisfaction des études	Dépression = Non (Effectifs)	Dépression = Oui (Effectifs)	Dépression = Non (Fréquences)	Dépression = Oui (Fréquences)
0.0	4	6	0.000143	0.000215
1.0	1593	3856	0.057101	0.138218
2.0	2070	3768	0.074199	0.135063
3.0	2467	3353	0.088429	0.120188
4.0	3095	3264	0.110940	0.116998
5.0	2334	2088	0.083662	0.074844

Tableau 9 Tableau de contingence

"ETAPE 1 : Test de comparaison graphique deux sous-étudiants"

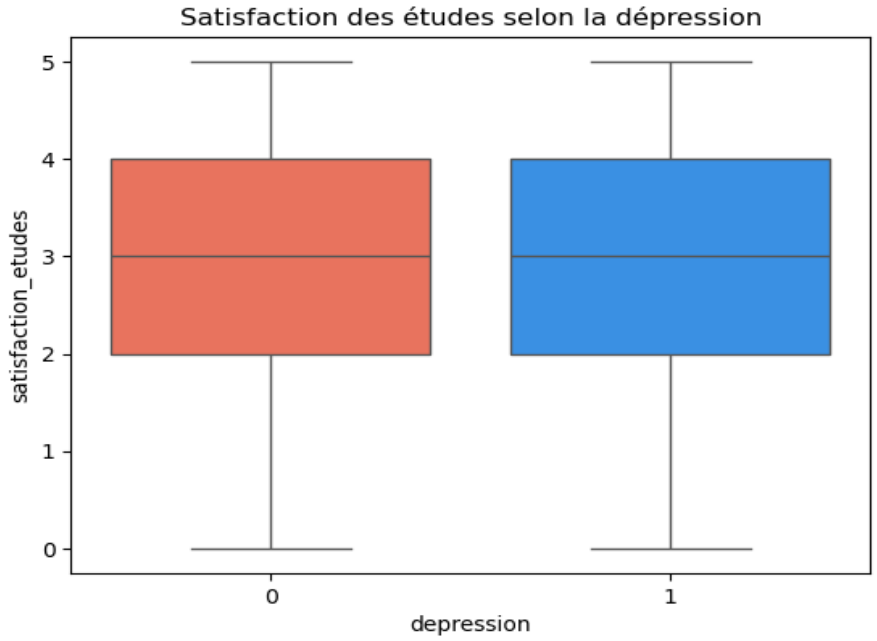


Figure 12Test de comparaison graphique

Observations graphiques :

- **Médiane** : La médiane de la satisfaction des études est identique pour les deux groupes, se situant autour de 3. Cela suggère que, globalement, les étudiants souffrant de dépression et ceux n'en souffrant pas ont une satisfaction académique médiane similaire.
- **Distribution** : Les quartiles des deux groupes couvrent des plages similaires, ce qui indique une répartition comparable des scores de satisfaction entre les étudiants souffrant de dépression et ceux n'en souffrant pas. Cette similarité dans la distribution des scores renforce l'idée qu'il n'y a pas de différence marquée entre les deux groupes en termes de satisfaction.
- **Étendues** : Les valeurs minimales et maximales sont identiques pour les deux groupes, ce qui suggère une dispersion comparable des scores. Les deux groupes présentent des niveaux de satisfaction qui varient de manière similaire, sans différence notable dans les extrêmes des données.
- **Symétrie et Outliers** : Aucune des distributions ne présente de valeurs aberrantes, et les deux groupes semblent relativement symétriques. L'absence de valeurs extrêmes et de déviations importantes indique que les données sont équilibrées et que la distribution des scores de satisfaction est homogène dans les deux groupes.

''ETAPE 2 : Estimer les statistiques de base''

depression	count	mean	std	min	25%	50%	75%	max
0 (sans dépression)	11,565	3.22	1.33	0.0	2.0	3.0	4.0	5.0
1 (avec dépression)	16,336	2.75	1.35	0.0	2.0	3.0	4.0	5.0

Tableau 10 Estimation de base

La **moyenne de satisfaction** est plus faible pour les étudiants souffrant de dépression (2.75) par rapport à ceux qui n'en souffrent pas (3.22), ce qui suggère que la dépression pourrait être associée à une perception moins positive de leur expérience académique.

Les **dispersion et quartiles** sont similaires entre les deux groupes, ce qui montre que, bien que les moyennes diffèrent, la répartition des scores de satisfaction est comparable dans les deux groupes.

Il est important de noter que les **valeurs extrêmes** (0.0 et 5.0) existent dans les deux groupes, ce qui signifie qu'il y a une variabilité importante dans les réponses des étudiants, même au sein de chaque groupe.

''ETAPE 3 : Tester la normalité des données dans chaque sous-population''

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

Donc pour notre étude : soit l'hypothèse suivante

H_0 : La distribution de satisfaction des études chez les étudiants souffrant de dépression suit la loi normale.

$H1$: La distribution de satisfaction des études chez les étudiants ne souffrant pas de dépression ne suit pas la loi normales.

Test de normalité de Shapiro	
Satisfaction étude	P-value
Étudiants souffrant de dépression :	**6.177610170128729e-74**
Étudiants sans dépression :	**8.085526933724967e-66**

Tableau 11 Test de normalité shapiro

Le résultat du test révèle **des p-values < 0.05** pour les deux types d'etudiants donc on rejette $H0$. Par conséquent nous ne pouvons utiliser **le test de Student** mais plutôt un test non paramétrique comme le **test de Kruskal-Wallis** ou le **test de Wilcoxon** pour tester l'égalité des moyennes de satisfaction des études chez les etudiant souffrant de depression ou pas.

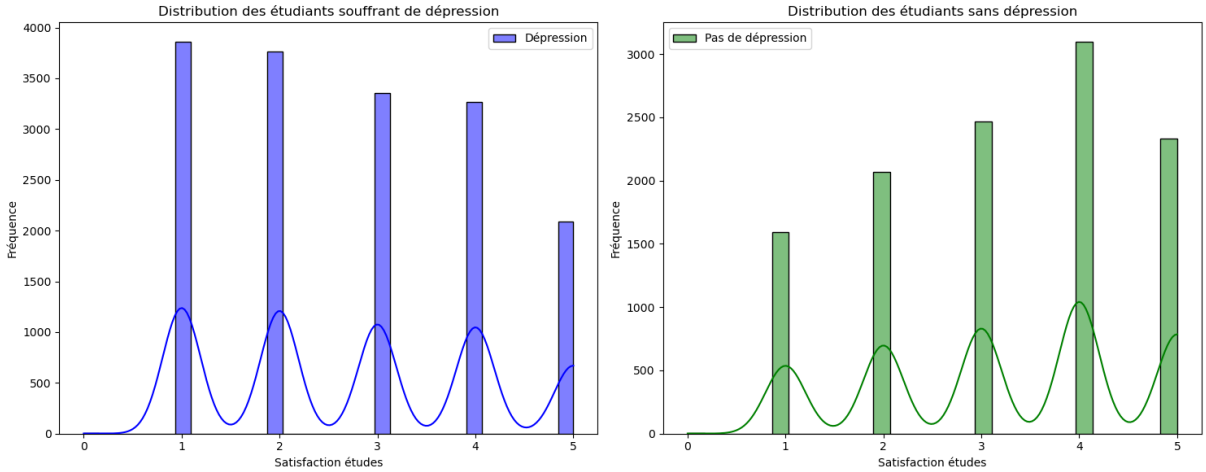


Figure 13 Graphique de normalité

ETAPE 4 : Tester l'égalité des variances (pas obliger vu que sa condition d'utilisation est respectée **à savoir la normalité).**

HYPOTHESE :

H_0 : σ depressifs = σ non-depressifs

$H1$: σ depressifs # µnon-depressifs

Test de levene pour l'égalité des variances :
leveneResult(statistic=13.159939601623353, pvalue=value = 0.00028650812611718136

Tableau 12 Egalité de la variance

Le résultat du test révèle une **p-value= 0.00028 < 0,05** on rejeter $H0$

ETAPE 5 : Tester l'égalité des moyennes : Test de Mann-Whitney

Nb : nous avons eu le choix entre le test de wilcoxon, le test de kruskal-walis et le test de Mann-Whitney, et notre regard s'est porté sur le test de de Mann-Whitney.

Pourquoi ?

Critère	Test de Wilcoxon	Test de Mann-Whitney	Test de Kruskal-Wallis
Type de comparaison	**Deux échantillons appariés ou dépendants**	**Deux échantillons indépendants**	**Trois groupes ou plus indépendants**
Hypothèse	Compare les médianes des paires appariées	Compare les distributions des deux groupes	Compare les distributions de trois groupes ou plus
Type de données	Ordinales ou continues (avec différences appariées)	Ordinales ou continues non normalement distribuées	Ordinales ou continues non normalement distribuées
Exemple d'application	Mesures avant-après sur les mêmes individus	Comparaison de deux groupes indépendants	Comparaison de plusieurs groupes indépendants
Test statistique	Test de Wilcoxon pour échantillons appariés	Test U de Mann-Whitney	Test de Kruskal-Wallis

HYPOTHESE :

H_0 : μ dépressifs = μ moyenne non-dépressifs

$H1$: μ moyenne dépressifs # μ moyenne non-dépressifs

Test de Mann-Whitney pour l'égalité des moyennes
Test de Mann-Whitney **Result (Statistic: 76236987.0, pvalue= 1.3620060777186873e-173)**

Tableau 13Test de Mann-Whitney pour l'égalité des moyennes

Le résultat du test donne une p-value= $\mathbf{1.36 \times 10^{-173} < 0.05}$ on peut donc rejeter $H0$ on conclut : qu'il il existe une **différence statistiquement significative** entre les deux groupes en termes de **distributions de satisfaction des études**. En d'autres termes, les deux groupes diffèrent de manière importante dans leur satisfaction des études.

2) **Test d'égalité de moyennes** : *Les niveaux de satisfaction au travail diffèrent-ils significativement selon le diplôme suivi ?*

Dans ce cas on veut déterminer si le type de diplôme (par exemple, un Bac, un Bac+2, un Bac+5, etc.) influence de manière significative le **niveau de satisfaction au travail**. Est-ce que les personnes ayant un diplôme spécifique sont plus ou moins satisfaites de leur travail que celles ayant d'autres diplômes ?

De ce fait nous procèderons à une série de test tout en respectant les étapes

"ETAPE 1 : Test de comparaison graphique deux sous-population"

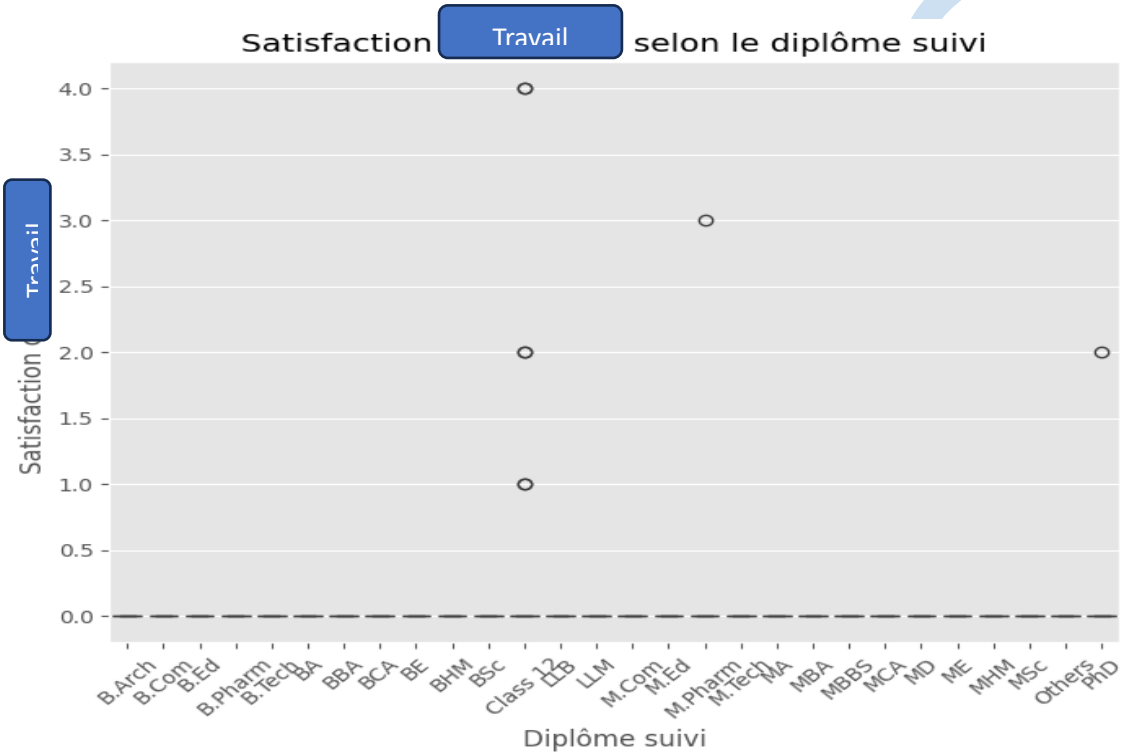


Figure 14 Satisfaction travail selon le diplôme suivi

"ETAPE 2 : Estimer les statistiques de base

Diplôme	Moyenne Satisfaction	Écart Type Diplôme_suivi
B.Arch	0.000000	1.354627
B.Com	0.000000	1.313457
B.Ed	0.000000	1.323866
B.Pharm	0.000000	1.351468
B.Tech	0.000000	1.323777
BA	0.000000	1.394479
BBA	0.000000	1.313250
BCA	0.000000	1.347611
BE	0.000000	1.315736
BHM	0.000000	1.354182
BSc	0.000000	1.398316
Class 12	0.002303	1.376500
LLB	0.000000	1.319849
LLM	0.000000	1.341006
M.Com	0.000000	1.363415
M.Ed	0.000000	1.357633
M.Pharm	0.005155	1.337204
M.Tech	0.000000	1.363793
MA	0.000000	1.360182
MBA	0.000000	1.393077
MBBS	0.000000	1.418580
MCA	0.000000	1.376713
MD	0.000000	1.354321
ME	0.000000	1.327929
MHM	0.000000	1.370657
MSc	0.000000	1.381537
Others	0.000000	1.336621
PhD	0.003831	1.377108

Tableau 14Estimation statistique

ETAPE 3 : Tester la normalité des données dans chaque sous-population

$H0$: la distribution suit une loi normale

$H1$: la distribution ne suit pas une loi normale

Diplôme	Valeur
B.Arch	None
B.Com	None
B.Ed	None
B.Pharm	None
B.Tech	None
BA	None
BBA	None
BCA	None
BE	None
BHM	None
BSc	None
Class 12	1.3802222133168593e-95
LLB	None
LLM	None
M.Com	None
M.Ed	None
M.Pharm	4.149120889566308976e-47
M.Tech	None
MA	None
MBA	None
MBBS	None
MCA	None
MD	None
ME	None
MHM	None
MSc	None
Others	None
PhD	3.4816983021577316e-45

ETAPE 4 : Tester l'égalité des variances

Test de Levene pour l'égalité des variances (plus robuste)

Test de Levene pour l'égalité des moyennes
Levene's test statistic: 0.7969946602044147, **P-value: 0.7613407525182427)**

ETAPE 5 : Tester l'égalité des moyennes : Test de Kruskal-Wallis

Test de Kruskal-Wallis
Kruskal-Walli's statistic: 25.32725578247819 **P-value**: 0.5561277145169576

V) TEST D'INDÉPENDANCE DEUX VARIABLES QUALITATIVES

L'objectif de cette partie est de répondre au travers de comparaison de tests de liaison à certaines questions suscitant un intérêt général

1) La dépression est-elle indépendante des habitudes alimentaires ?

Nous répondrons à cette question en suivant certaines étapes

"ETAPE 1 : Test de comparaison graphique

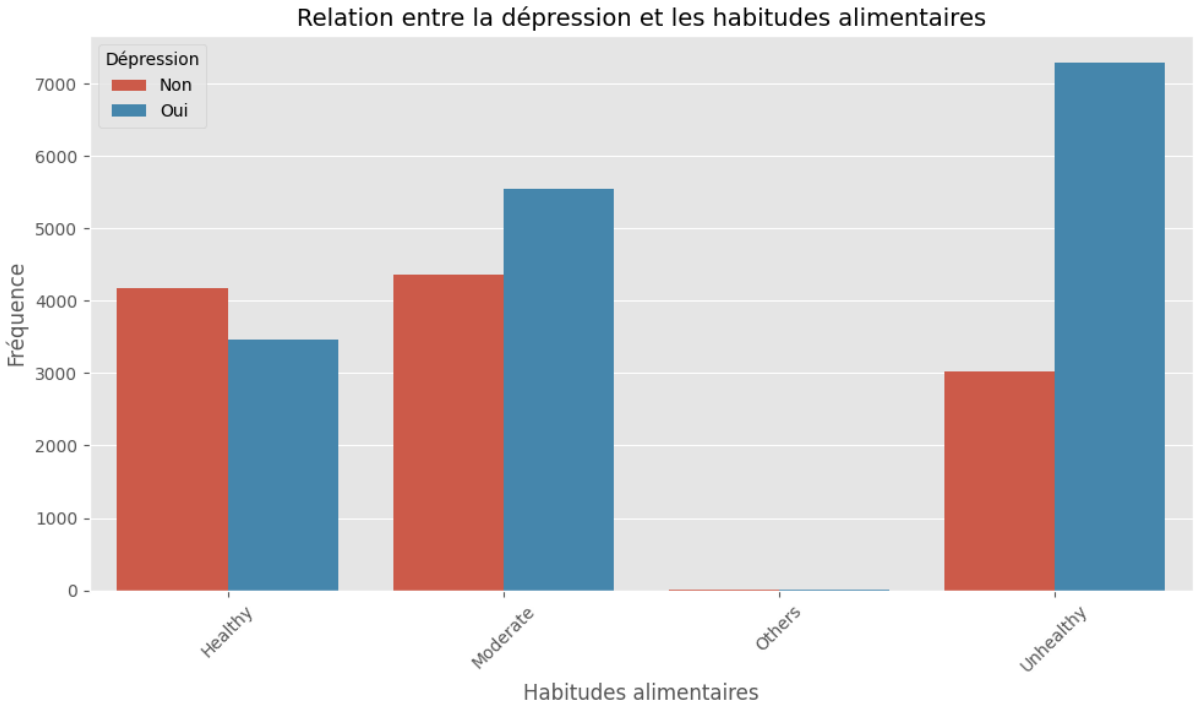


Figure 15 Relation entre dépression et habitudes alimentaire

"ETAPE 2 Vérification de la condition de Cochran

La condition de Cochran fait référence à la nécessité d'avoir des tailles d'échantillons adéquates et une distribution des variances assez homogène pour pouvoir effectuer des tests d'égalité des variances tels que le test de Bartlett de manière fiable. Si cette condition n'est pas remplie, il est souvent recommandé d'utiliser des tests alternatifs plus robustes

Habitudes Alimentaires	Healthy	Moderate	Others	Unhealthy
Non	4177	4363	4	3019
Oui	3472	5558	8	7297

Effectifs théoriques :

[[3.17031282e+03 4.11199810e+03 4.97368987e+00 4.27571539e+03]

[4.47868718e+03 5.80900190e+03 7.02631013e+00 6.04028461e+03]]

La condition de Cochran est respectée : True

"ETAPE 3 Test de liaison de Khi-2

Le test de liaison de Khi-2 permet de déterminer si la répartition des données dans un tableau de contingence (c'est-à-dire la manière dont les fréquences observées se répartissent dans les différentes catégories) est compatible avec l'hypothèse d'indépendance entre les variables.

Test de liaison de Khi-2
Statistique du chi-carré : 1203.26724929502
P-value : 1.4332378809995893e-260

Effectifs théoriques :

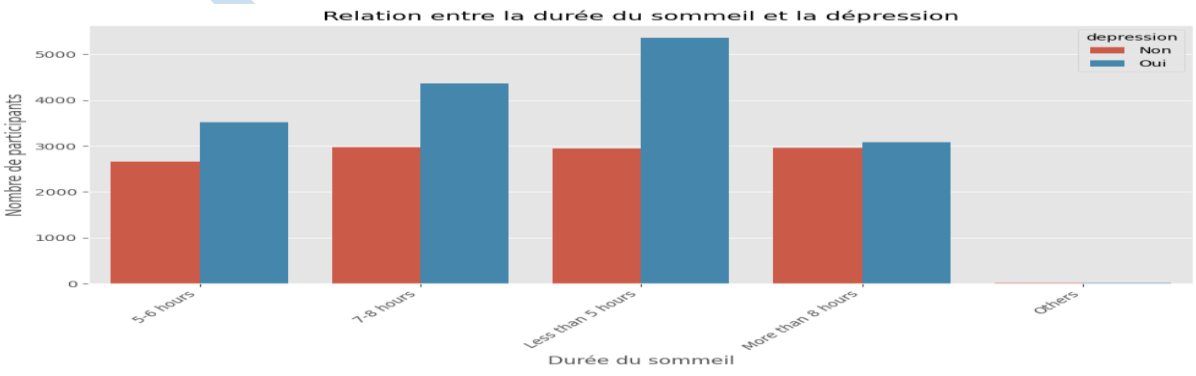
[[3.17031282e+03 4.11199810e+03 4.97368987e+00 4.27571539e+03]

[4.47868718e+03 5.80900190e+03 7.02631013e+00 6.04028461e+03]]

Le test est significatif, il y a une association entre 'depression' et 'habitudes_alimentaires'.

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

"ETAPE 1 : Test de comparaison graphique



"ETAPE 2 Vérification de la condition de Cochran

Durée de Sommeil	5-6 hours	7-8 hours	Less than 5 hours	More than 8 hours	Others
Non	2665	2975	2948	2966	9
Oui	3516	4371	5361	3078	9

Effectifs théoriques :

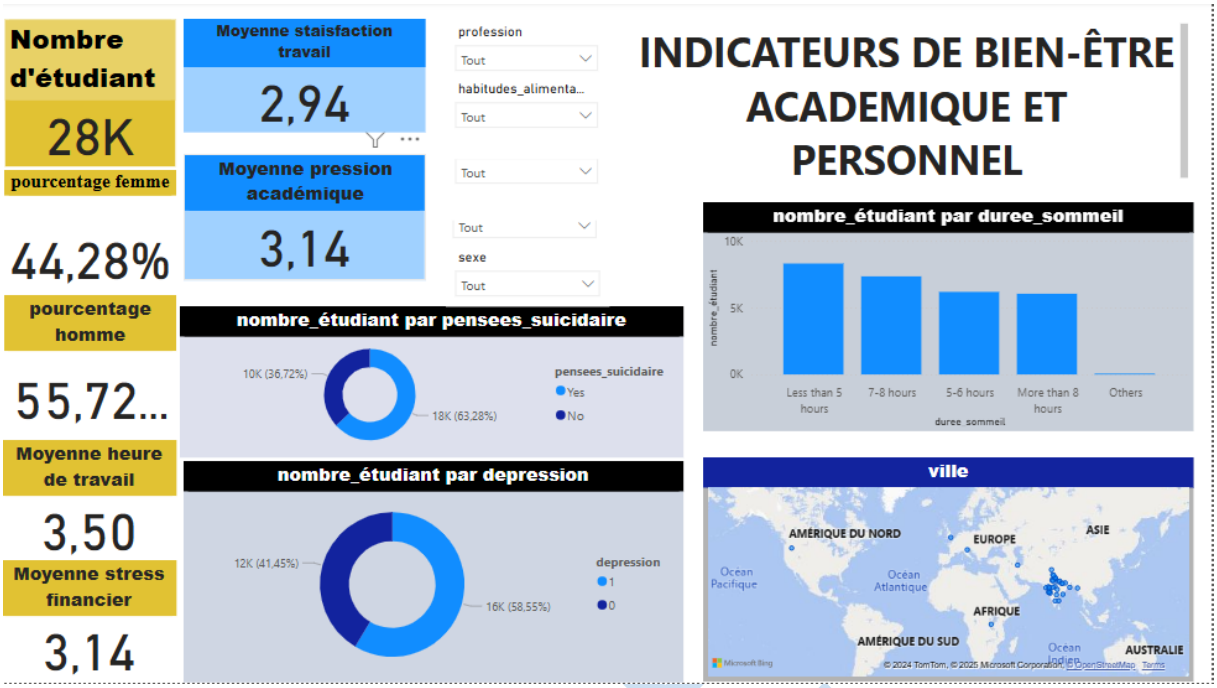
277.13495443548004

"ETAPE 3 Test de liaison de Khi-2

Test de liaison de Khi-2
Chi2ContingencyResult (statistic=277.13495443548004, **pvalue**=9.240458086403438e-59
expected_freq=array ([[2561.86475733, 3044.7271489, 3443.86576099, 2505.08179798, 7.46053481], [3619.13524267, 4301.2728511, 4865.13423901, 3538.91820202, 10.53946519]]))

La durée du sommeil est significativement liée à la dépression

TABLEAU DE BORD



CONCLUSION

L'objectif de cette analyse était d'explorer et de comprendre les caractéristiques des données relatives à des variables liées à la santé mentale, la satisfaction académique et professionnelle, et les habitudes de vie des étudiants. L'objectif était d'identifier des tendances, des relations entre les variables, ainsi que d'analyser la distribution de ces variables afin de mieux orienter les futures analyses.

L'analyse exploratoire a révélé plusieurs points intéressants concernant les distributions des variables. Les variables telles que **pression_liee_au_travail** et **satisfaction_travail** présentent une asymétrie très marquée, indiquant une concentration de faibles valeurs et une longue queue vers des valeurs extrêmes. **age**, **pression_academique**, et **stress_financier** ont une faible asymétrie, indiquant une distribution relativement symétrique avec de petites variations. Les variables **nombre_heure_travail_etude** et **moyenne_notes** présentent également une légère asymétrie, mais ces valeurs restent assez équilibrées. Enfin, les variables liées à la satisfaction des études et au stress financier montrent une faible asymétrie, suggérant une distribution relativement homogène parmi la population.

Afin d'améliorer l'analyse, plusieurs recommandations peuvent être faites. D'abord, il est recommandé de normaliser ou de transformer les variables fortement asymétriques (par exemple, **pression_liee_au_travail** et **satisfaction_travail**) afin de rendre leur distribution plus proche d'une distribution normale, ce qui pourrait améliorer la performance de certains modèles d'analyse. De plus, il serait pertinent d'explorer les causes sous-jacentes des fortes asymétries dans certaines variables, pour mieux comprendre les facteurs influençant la pression académique et la satisfaction au travail.

Cependant, cette analyse présente certaines limites. Les données manquantes ou incomplètes n'ont pas été prises en compte, ce qui pourrait affecter la précision des résultats. De plus, l'analyse s'est concentrée sur un sous-ensemble de variables, sans intégrer d'autres facteurs (sociaux, économiques, environnementaux) pouvant également influencer la santé mentale, le stress et la satisfaction des étudiants. L'asymétrie extrême de certaines variables pourrait également limiter la généralisation des résultats si elles ne sont pas correctement traitées.

Pour améliorer les résultats, des analyses multivariées pourraient être envisagées afin de comprendre les relations complexes entre les variables et leur influence sur la santé mentale des étudiants. Une analyse temporelle permettrait également d'observer l'évolution de ces variables au fil du temps et de vérifier si des événements externes affectent ces variables. Enfin, la collecte de données supplémentaires sur des facteurs sociaux, familiaux et économiques pourrait offrir une vue plus complète et enrichir l'analyse pour mieux comprendre les dynamiques sous-jacentes.

Ainsi, bien que cette analyse ait fourni des premiers insights intéressants, des recherches plus approfondies sont nécessaires pour mieux comprendre les causes des disparités observées et mettre en place des stratégies d'intervention ciblées.

Annexe

1. Bibliographie

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

2. Source du code Python

```
# STATISTIQUE INFERETIELLE

# IMPORTATION DES BIBLIOTHEQUES NECESSAIRES

"Bibliotheque principale"
import pandas as pd
import numpy as np
from scipy import stats
import random
import warnings
import pingouin as pg

"Bibliotheque de visualisation"
import matplotlib.pyplot as plt
from matplotlib.colors import LinearSegmentedColormap
import seaborn as sns
import plotly.graph_objects as go
import plotly.express as px
import squarify
%matplotlib inline
"Definir une graine aleatoire"
rs = 42

"Ignorer les avertissement"
warnings.filterwarnings("ignore")
```

```
"Definir la pallete de couleur pour seaborn"

colors= ['#1c76b6', '#a7dae9', '#eb6a20', '#f59d3d', '#677fa0',
'#d6e4ed', '#f7e9e5']

sns.set_palette(colors)


##########################################################
##################################

            "# PREPARATION DES DONNEES "

##########################################################
##################################


# 1) Imporattion du jeu de données


SD_train =
pd.read_csv('C:/users/yoboh/OneDrive/Bureau/INSSEDS/MINI
PROJET/PROJETS/STAT_INFERENCE/DATA
bases/Student_Depression.csv', sep=';')

SD_test =
pd.read_csv('C:/users/yoboh/OneDrive/Bureau/INSSEDS/MINI
PROJET/PROJETS/STAT_INFERENCE/DATA
bases/Student_Depression.csv', sep=';')


print("First 5 rows of our dataset:")

SD_train.head()

SD_train.tail()


#*inoformation sur le jeu de données*


print(f"Il y a {SD_train.shape[1] }colonnes et {SD_train.shape[0]}
lignes dans l'ensemble de données du train.")


print("Noms des colonnes et type de données de chaque colonne:")

SD_train.dtypes
```

```
print("Il y a {} doublons dans le jeu de
données.".format(SD_train.duplicated().sum()))


# visualisation des valeurs manquantes

print("Vérification des valeurs manquantes dans chaque colonne:")

print(SD_train.isnull().sum())


#1) affichache graphique des valeur manquantes

plt.figure(figsize=(18,12))

plt.title("Visualisation des valeurs manquantes")

sns.heatmap(SD_train.isnull(), cbar=False,
cmap=sns.color_palette(colors), yticklabels=False);

plt.show()



# Arrangement de certains element du jeu de doonnée


# Save 'id' column for submission

test_ids = SD_test['id']


# Supprimer la colonne « id » dans les deux ensembles de données

SD_train = SD_train.drop(['id'], axis=1)

SD_test = SD_test.drop(['id'], axis=1)


# Define la colonne cible

target_column = 'depression'


# Select categorical and numerical columns (initial)

categorical_columns =
SD_train.select_dtypes(include=['object']).columns

numerical_columns =
SD_train.select_dtypes(exclude=['object']).columns.drop(target_colum
n)
```

```
# Print out column information
print("Target Column:", target_column)
print("\nCategorical Columns:", categorical_columns.tolist())
print("\nNumerical Columns:", numerical_columns.tolist())


# presentation du nombre de chaque element
for column in categorical_columns:
    num_unique = SD_train[column].nunique()
    print(f"'{column}' has {num_unique} unique categories.")


# Imprimez les 10 premières valeurs uniques pour chaque colonne catégorielle
for column in categorical_columns:
    print(f"\nTop value counts in '{column}':\n{SD_train[column].value_counts().head(10)}")



##########################################################################
                "# ANALYYSES EXPLORATOIRE DES DONNEES"
##########################################################################


 "#  1) Analyse univariee"


# affichage des parametre statistiques
print("The skewness of columns:")
print(SD_train[numerical_columns].skew())



# a) Distribution des variables numériques
```

```
numerical_columns_to_plot = ["age", "moyenne_notes",
"nombre_heure_travail_etude"]


plt.figure(figsize=(12, 16))


for i, column in enumerate(numerical_columns_to_plot):
    plt.subplot(3, 1, i+1)
    sns.histplot(data=SD_train, x=column, kde=False, bins=20,
color=colors[0])
    plt.title(f'Distribution of {column}')
    sns.despine()


plt.tight_layout()
plt.show()


# b) Distribution des caractéristiques catégorielles
# Liste des colonnes à afficher
categorical_columns_to_plot = ['sexe', 'profession',
'pression_academique',
                               'pression_liee_au_travail', 'satisfaction_etudes',
                               'pensees_suicidaire', 'stress_financier',
                               'antecedants_familiaux_maladie_mentale']


# Plot countplots pour chaque colonne catégorielle
for column in categorical_columns_to_plot:


    plt.figure(figsize=(8, 5))


    if column == 'stress_financier':
        sns.countplot(data=SD_train, x=column, palette='Set2')
    elif column == 'satisfaction_etudes':
        sns.countplot(data=SD_train, x=column, palette='Set2')
```

```
    elif column == 'pression_academique':
        sns.countplot(data=SD_train, x=column, palette='Set2')
    elif column == 'pression_liee_au_travail':
        sns.countplot(data=SD_train, x=column, palette='Set2')
    elif column == 'profession':
        sns.countplot(data=SD_train, x=column, palette='Set2')
    elif column == 'pensees_suicidaire':
        sns.countplot(data=SD_train, x=column, palette=['lightgreen',
'lightyellow'])
    elif column == 'sexe':
        sns.countplot(data=SD_train, x=column, palette=['lightgreen',
'lightyellow'])
    elif column == 'antecedants_familiaux_maladie_mentale':
        sns.countplot(data=SD_train, x=column, palette=['lightgreen',
'lightyellow'])
        sns.countplot(data=SD_train, x=column)

    plt.title(f'Countplot of {column}')
    plt.tight_layout()
    plt.show()


 "#  2) Analyse Bivariee"

# 2.1) Distribution des caractéristiques numériques selon la présence de
dépression

bi_palette = [colors[3], colors[0]]

for column in numerical_columns_to_plot:
    plt.figure(figsize=(8, 6))
    sns.violinplot(data=SD_train, x=target_column, y=column,
palette=bi_palette)
```

```
    plt.title(f'Distribution of {column} by Target')


    plt.tight_layout()
    plt.show()


# 2.2) Relation entre les colonnes catégorielles et une variable cible


cmap = LinearSegmentedColormap.from_list("custom_cmap",
bi_palette)


# Parcourez chaque colonne catégorielle de votre liste
for column in categorical_columns_to_plot:
   # Graphique à barres empilées
    pd.crosstab(SD_train[column],
SD_train[target_column]).plot(kind='bar', stacked=True,
colormap=cmap, figsize=(8, 6))
    plt.title(f"Stacked Bar Plot of {column} and Target")
    plt.xlabel(column)
    plt.ylabel("Frequency")
    plt.show()


# 2.3 Découverte des professions
# Calculate frequencies
# Liste des professions que tu souhaites inclure
professions = [
    "Student", "Architect", "Teacher", "Digital Marketer",
    "Content Writer", "Chef", "Doctor", "Pharmacist",
    "Civil Engineer", "UX/UI Designer"
]


# Valeurs fictives associées à chaque profession (pour l'exemple)
# Adapte les valeurs en fonction de tes données réelles si nécessaire
```

```
values = [500, 40, 40, 25, 15, 50, 70, 35, 45, 20]  # Exemple de fréquence ou taille


# Choisir des couleurs pour les professions (optionnel)
colors = plt.cm.Paired(range(len(professions)))


# Tracer le treemap
plt.figure(figsize=(12, 8))
squarify.plot(sizes=values, label=professions, color=colors, pad=True)
plt.title("Treemap des Professions")
plt.axis("off")  # Désactiver les axes pour une présentation plus claire
plt.show()


# Deuxieme partie des professions
# Calculate frequencies
value_counts = SD_train['profession'].value_counts()
sizes = value_counts.values[:20]  # Show only the top 20 for readability


plt.figure(figsize=(12, 8))
squarify.plot(sizes=sizes, label=value_counts.index[:20], color=colors, pad=True)
plt.title(f"Treemap of profession (Top 20)")
plt.axis("off")
plt.show()



# 2.4 Create a DataFrame for the top 20 professions and their relationship with depression
top_n_professions = 10
profession_counts = SD_train['profession'].value_counts().nlargest(top_n_professions)
filtered_data = SD_train[SD_train['profession'].isin(profession_counts.index)]
```

```
# Create a summary DataFrame
sankey_data = filtered_data.groupby(['profession',
'depression']).size().reset_index(name='Count')


# Define the source and target for the Sankey chart
labels = list(sankey_data['profession'].unique()) + ['No depression',
'depression']
source_indices = []
target_indices = []


for _, row in sankey_data.iterrows():
    profession_index = labels.index(row['profession'])
    depression_index = labels.index('depression' if row['depression'] ==
1 else 'No depression')
    source_indices.append(profession_index)
    target_indices.append(depression_index)


# Create a Sankey chart
fig = go.Figure(data=[go.Sankey(
    node=dict(
        pad=15,
        thickness=20,
        line=dict(color="black", width=0.5),
        label=labels,
        color='blue'
    ),
    link=dict(
        source=source_indices,  # Indices correspond to labels, e.g., A=0,
B=1, C=2...
        target=target_indices,
        value=sankey_data['Count']
    )
```

```
)])

fig.update_layout(
    title_text="Sankey Diagram of Profession and Depression",
    font_size=10,
    width=700,
    height=600)

fig.show(renderer="browser")

# 2.5 Les diplomes suivis
# Calculate frequencies
value_counts = SD_train['diplome_suivi'].value_counts()
sizes = value_counts.values[:20]  # Show only the top 20 for readability

plt.figure(figsize=(12, 8))
squarify.plot(sizes=sizes, label=value_counts.index[:20], color=colors,
pad=True)
plt.title(f"Treemap of diplome_suivi (Top 20)")
plt.axis("off")
plt.show()


# 2.6  tableau des 10 meilleurs diplome suivi et depression
# Get the top 10 most common professions
top_professions =
SD_train['diplome_suivi'].value_counts().nlargest(10).index

# Filter the DataFrame for the top 10 professions
filtered_df = SD_train[SD_train['diplome_suivi'].isin(top_professions)]

# Create a new DataFrame for aggregation
```

```
agg_data = filtered_df.groupby(['diplome_suivi',
'depression']).size().reset_index(name='Count')


# Create a sunburst chart
fig = px.sunburst(agg_data,
                  path=['diplome_suivi', 'depression'],
                  values='Count',
                  title='Sunburst Chart of Top 10 Degrees and depression',
                  color='Count',
                  color_continuous_scale=px.colors.sequential.Oranges[:])


# Show the figure
fig.show(renderer="browser")



#2.7 distribution sur la variable depression


# Calculate counts for the pie chart and add labels
class_counts = SD_train[target_column].value_counts().sort_index()
labels = ["No depression", "depression"]
plt.figure(figsize=(6, 6))
plt.pie(class_counts, colors=bi_palette, labels=labels,
    autopct='%1.1f%%', startangle=90)
plt.title('Distribution of Target Variable')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a
circle.


plt.show()


# 2.8 correlation entre les variables
# Calculate the correlation matrix
#correlation entre les variables
```

```
# Heatmap avec seaborn (statique)
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='RdYlBu', fmt='.2f')
plt.title('Heatmap of Correlation Matrix')
plt.show()

# Calculate the correlation matrix
correlation_matrix = SD_train.corr(numeric_only=True)

# Create an interactive heatmap with orange and blue colors
fig = px.imshow(correlation_matrix,
                text_auto=True,  # Display correlation values
                color_continuous_scale='RdYlBu',  # Color scale with shades
of red, yellow, and blue
                title='Heatmap of Correlation Matrix',
                aspect='auto')

# Show the plot
fig.show(renderer="browser")



# 2.9
# Vérification des données
print(SD_train[['age', 'pression_liee_au_travail',
'depression']].isnull().sum())

# Nettoyer les données
SD_train_copy = SD_train.dropna(subset=['age',
'pression_liee_au_travail', 'depression'])

# Créer des bins pour l'âge et la pression liée au travail
SD_train_copy['age_bin'] = pd.cut(SD_train_copy['age'],
bins=10).astype(str)
```

```
SD_train_copy['pression_liee_au_travail_bin'] =
pd.cut(SD_train_copy['pression_liee_au_travail'], bins=10).astype(str)


# Vérification des bins

print(SD_train_copy[['age_bin',
'pression_liee_au_travail_bin']].head())


# Créer une table pivotée

heatmap_data = SD_train_copy.pivot_table(index='age_bin',
columns='pression_liee_au_travail_bin', values='depression',
aggfunc='mean')


# Vérification de la table pivotée

print(heatmap_data)


# Créer la heatmap interactive

fig = px.imshow(heatmap_data.values,  # Utilisation uniquement de la
matrice numérique

            labels=dict(x="pression_liee_au_travail Bin", y="age Bin",
color="depression"),

            text_auto=True,  # Afficher les valeurs de corrélation

            color_continuous_scale='RdYlBu',  # Palette de couleurs

            title='Heatmap of depression by age and Work Pressure',

            aspect='auto')


# Définir les labels de l'axe X et Y

fig.update_layout(

   yaxis=dict(

      tickmode='array',

      tickvals=list(range(len(heatmap_data.index))),

      ticktext=heatmap_data.index.astype(str).tolist()

   ),

   xaxis=dict(
```

```
        tickmode='array',
        tickvals=list(range(len(heatmap_data.columns))),
        ticktext=heatmap_data.columns.astype(str).tolist()
    )
)

# Forcer l'affichage dans le navigateur si nécessaire
fig.show(renderer="browser")


##########################################################
##################################
                " ANALYSE INFERENTIELLE ET PROCEDURE "
##########################################################
##################################

import pandas as pd
import scipy.stats as stats
import numpy as np
henoc =
pd.read_csv('C:/users/yoboh/OneDrive/Bureau/INSSEDS/MINI
PROJET/PROJETS/STAT_INFERENCE/DATA
bases/Student_Depression.csv', sep=';')

I) ESTIMATION STATISTIQUE
  1) "intervalle de confiance "  : proportion d'étudiants ayant déjà eu des
pensées suicidaires

# 1. Intervalle de confiance pour la proportion d'étudiants ayant eu des
pensées suicidaires
suicidal_thoughts =
henoc['pensees_suicidaire'].value_counts(normalize=True)
proportion = suicidal_thoughts['Yes']  # Calculer la proportion de 'Oui'

# Intervalle de confiance à 95% pour la proportion
```

```
z = 1.96  # Valeur critique pour un intervalle de confiance à 95%
n = len(henoc)  # Nombre total d'étudiants
ic_lower = proportion - z * np.sqrt(proportion * (1 - proportion) / n)
ic_upper = proportion + z * np.sqrt(proportion * (1 - proportion) / n)

print(f"Intervalle de confiance pour la proportion d'étudiants ayant eu des pensées suicidaires : ({ic_lower:.4f}, {ic_upper:.4f})")
```

2) " MOYENNE et de la médiane ":

2_1) heures de travail ou d'études pour les étudiants souffrant de dépression

```
# Moyenne et médiane des heures de travail pour les étudiants dépressifs
depressed_students = henoc[henoc['depression'] == 1]['nombre_heure_travail_etude']
mean_hours_depressed = depressed_students.mean()
median_hours_depressed = depressed_students.median()

print(f"Moyenne des heures de travail/études pour les étudiants dépressifs : {mean_hours_depressed:.2f} heures")
print(f"Médiane des heures de travail/études pour les étudiants dépressifs : {median_hours_depressed:.2f} heures")
```

2.2) # Stress financier pour les étudiants avec et sans dépression

```
stress_depressed = henoc[henoc['depression'] == 1]['stress_financier']
stress_not_depressed = henoc[henoc['depression'] == 0]['stress_financier']

mean_stress_depressed = stress_depressed.mean()
median_stress_depressed = stress_depressed.median()

mean_stress_not_depressed = stress_not_depressed.mean()
```

```
median_stress_not_depressed = stress_not_depressed.median()


print(f"Moyenne du stress financier pour les étudiants dépressifs : {mean_stress_depressed:.2f}")

print(f"Médiane du stress financier pour les étudiants dépressifs : {median_stress_depressed:.2f}")


print(f"Moyenne du stress financier pour les étudiants non dépressifs : {mean_stress_not_depressed:.2f}")

print(f"Médiane du stress financier pour les étudiants non dépressifs : {median_stress_not_depressed:.2f}")
```

II. TEST DE COMPARAISON DE POPULATIONS (TEST D'ÉGALITÉ DE MOYENNES)

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

```
##### "ETAPE 1 : Comparer graphiquement les deux sous-populations"
```

Nous allons comparer graphiquement la satisfaction des études entre les étudiants souffrant de dépression (variable : depression) et ceux n'en souffrant pas.

```
import seaborn as sns

import matplotlib.pyplot as plt


sns.boxplot(x='depression', y='satisfaction_etudes', data=henoc, palette=["#FF6347", "#1E90FF"])

plt.title("Satisfaction des études selon la dépression")

plt.show()


#### ETAPE 2 : Estimer les statistiques de base

henoc.groupby('depression')['satisfaction_etudes'].describe()
```

ETAPE 3 : Tester la normalité des données dans chaque sous-population

Nous devons tester si les données suivent une distribution normale dans chaque sous-population. Nous utiliserons le test de Shapiro-Wilk pour tester la normalité.

```
from scipy.stats import shapiro


# Séparer les groupes

depression_group = henoc[henoc['depression'] ==
1]['satisfaction_etudes']

no_depression_group = henoc[henoc['depression'] ==
0]['satisfaction_etudes']


# Test de Shapiro-Wilk

stat_depression, p_value_depression = shapiro(depression_group)

stat_no_depression, p_value_no_depression =
shapiro(no_depression_group)


print("Test de normalité pour les étudiants souffrant de dépression: p-
value =", p_value_depression)

print("Test de normalité pour les étudiants sans dépression: p-value =",
p_value_no_depression)
```

ETAPE 4 : Tester l'égalité des variances

```
from scipy.stats import bartlett


# Test de Bartlett pour l'égalité des variances

statistic, p_value = bartlett(depression_group, no_depression_group)

print("Test d'égalité des variances pour la satisfaction des études
(Bartlett) : p-value =", p_value)

print(f"Statistic: {statistic}")
```

ETAPE 5 : Tester l'égalité des moyennes

Nous utiliserons un test t de Student si les conditions de normalité et d'égalité des variances sont remplies, sinon un test de Mann-Whitney pour comparer les moyennes entre les groupes.

```
from scipy.stats import mannwhitneyu

# Test de Mann-Whitney U pour des groupes indépendants
stat, p_value = mannwhitneyu(depression_group, no_depression_group)
print("Test de l'égalité des moyennes pour la satisfaction des études (Mann-Whitney U) : p-value =", p_value)
```

"ou"

```
from scipy.stats import ttest_ind

# Test t de Student ( prendre un test non parametrique )
stat, p_value = ttest_ind(depression_group, no_depression_group, equal_var=False)
print("Test de l'égalité des moyennes pour la satisfaction des études : p-value =", p_value)
```

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

#####"ETAPE 1 : Comparer graphiquement les deux sous-populations"

Nous allons comparer graphiquement la satisfaction au travail (variable : satisfaction_travail) selon les différents diplômes suivis (variable : diplome_suivi).

```
sns.boxplot(x='diplome_suivi', y='satisfaction_travail', data=henoc)
'diplome_suivi'
plt.title("Satisfaction au travail selon le diplôme suivi")
```

```
plt.xticks(rotation=90)
plt.show()


####    ETAPE 2 : Estimer les statistiques de base
henoc.groupby('diplome_suivi')['satisfaction_travail'].describe()


####  ETAPE 3 : Tester la normalité des données dans chaque sous-population


# Séparer les groupes
diploma_groups = henoc['diplome_suivi'].unique()


for diploma in diploma_groups:
   group_data = henoc[henoc['diplome_suivi'] == diploma]['satisfaction_travail']
   stat, p_value = shapiro(group_data)
   print(f"Test de normalité pour {diploma}: p-value = {p_value}")


####  ETAPE 4 : Tester l'égalité des variances


# Récupérer les groupes de diplômes
groups = [henoc[henoc['diplome_suivi'] == diploma]['satisfaction_travail'] for diploma in diploma_groups]


# Test de Levene
stat, p_value = levene(*groups)
print("Test d'égalité des variances pour la satisfaction au travail : p-value =", p_value)


####  ETAPE 5 : Tester l'égalité des moyennes
# Test t de Student pour chaque paire de groupes
from itertools import combinations
```

```
for group1, group2 in combinations(diploma_groups, 2):
    group1_data = henoc[henoc['diplome_suivi'] == group1]['satisfaction_travail']
    group2_data = henoc[henoc['diplome_suivi'] == group2]['satisfaction_travail']
    stat, p_value = ttest_ind(group1_data, group2_data)
    print(f"Test de l'égalité des moyennes entre {group1} et {group2}: p-value = {p_value}")
```

III. TEST D'INDÉPENDANCE DEUX VARIABLES QUALITATIVES

###1. Test de l'indépendance entre la dépression et les habitudes alimentaires (saines/modérées)

```
import pandas as pd
from scipy.stats import chi2_contingency

# Créez un tableau croisé des habitudes alimentaires et de la dépression
tableau_contingence = pd.crosstab(henoc['depression'], henoc['habitudes_alimentaires'])
print(tableau_contingence)

# Appliquez le test du chi carré
chi2, p_value, _, _ = chi2_contingency(tableau_contingence)

# Afficher le résultat
print("Test d'indépendance entre la dépression et les habitudes alimentaires : p-value =", p_value)
```

####2. Test de l'indépendance entre la durée du sommeil et la dépression

```
# Créez un tableau croisé de la durée du sommeil et de la dépression
tableau_contingence_sommeil = pd.crosstab(henoc['depression'], henoc['duree_sommeil'])
```

```
print(tableau_contingence_sommeil)

# Appliquez le test du chi carré

chi2_sommeil, p_value_sommeil, _, _ =
chi2_contingency(tableau_contingence_sommeil)


# Afficher le résultat

print("Test d'indépendance entre la durée du sommeil et la dépression :
p-value =", p_value_sommeil)
```