Student Depression Study - L3 MIAGE - 2024-2025

Marius Adjakotan, Nesrine Allouche, Tiana Rasolonzatovo, Olessya Miroshnichenko

2025-05-17

Table of contents

Introduction

La **santé mentale** des étudiants universitaires est une préoccupation croissante à l'échelle mondiale. Des études récentes indiquent qu'une part importante des jeunes adultes sont confrontés à des symptômes de dépression, liés à des facteurs variés : pression académique, incertitude financière, troubles du sommeil, ou encore problèmes personnels. Identifier précocement les étudiants à risque est crucial pour mettre en place des stratégies préventives et d'accompagnement.

Ce projet s'inscrit dans cette problématique : nous analysons un jeu de données regroupant plusieurs aspects de la vie étudiante afin de comprendre les déterminants de la dépression.

Notre objectif est de modéliser la probabilité de dépression à partir de données académiques, personnelles et sociales, dans le but d'identifier les leviers d'action prioritaires pour les établissements d'enseignement.

Context / Dataset Description

Pour ce projet, nous utilisons un dataset intitulé "Student Depression Dataset", disponible sur Kaggle. Ce dataset vise à explorer les facteurs susceptibles d'influencer la santé mentale des étudiants, en particulier la dépression. Les données ont été collectées via des questionnaires auto-administrés, où chaque étudiant a décrit sa situation personnelle, académique et sociale.

Le jeu de données couvre plusieurs dimensions importantes :

- Aspects académiques :
  - Moyenne académique (CGPA) : indicateur global des performances scolaires.
  - Pression académique : ressenti du stress lié aux études (examens, devoirs…).
  - Satisfaction dans les études : niveau de contentement vis-à-vis de la filière suivie.
- Mode de vie et bien-être :
  - Durée moyenne du sommeil.
  - Habitudes alimentaires.
  - Pression professionnelle (pour ceux qui travaillent en parallèle des études).
  - Nombre d'heures de travail/études par jour.
  - Satisfaction professionnelle (pour ceux ayant un emploi).
- Facteurs personnels :
  - Genre, Âge, Ville de résidence.
  - Stress financier.
  - Profession / Diplôme préparé.
  - Antécédents familiaux de troubles mentaux.
  - Pensées suicidaires (question binaire : oui/non).

La variable d'intérêt principale est la présence ou non d'une dépression déclarée par l'étudiant (variable binaire).

⊠ À noter :

Les données décrivent la situation des étudiants *au moment de l'enquête*, c'est-à-dire leur état de santé mentale et leurs conditions de vie *avant ou pendant* l'apparition d'une éventuelle dépression. Elles n'incluent pas les conséquences à long terme de la maladie ni de données cliniques approfondies.

Le dataset offre donc une photographie précieuse des *facteurs de risque potentiels* avant le début de symptômes dépressifs, ce qui permet d'identifier des profils à risque pour une détection précoce.

Il contient *27 901 enregistrements* et *18 variables*, chacun représentant un étudiant unique.

Description of the question

Notre question de recherche est formulée ainsi :

Quels sont les profils d'étudiants les plus vulnérables face à la dépression, et comment interagissent les facteurs académiques, sociaux et personnels dans la prédiction de ce risque ?

Cette question vise à :

- Identifier les **principaux facteurs prédictifs** de la dépression à partir des données collectées
- Explorer l'existence de **profils à risque** spécifiques (par exemple : étudiants cumulant stress académique élevé + mauvaise qualité de sommeil + difficultés financières)
- Analyser les **interactions potentielles** entre plusieurs facteurs pour comprendre comment ceux-ci se renforcent mutuellement dans l'augmentation du risque
- **Évaluer les disparités selon des sous-groupes :** par exemple en fonction du genre, de la ville, ou du niveau d'études, afin d'identifier des populations particulièrement vulnérables
- **Mettre en lumière des facteurs potentiellement "modifiables"** (ex : habitudes de sommeil, satisfaction, etc.) afin de proposer des pistes concrètes pour la prévention.
- **Identifier d'éventuelles incohérences ou paradoxes** (ex : des étudiants avec de bonnes notes mais malgré tout très dépressifs) pour enrichir la compréhension des dynamiques sous-jacentes

Ainsi, ce projet combine une approche descriptive (exploration des tendances globales et des disparités) et une approche prédictive.

Methodology

Data clean-up procedures

Avant de procéder à l'analyse, un nettoyage minutieux du jeu de données a été effectué afin d'assurer la qualité des résultats obtenus. Cela comprend les étapes suivantes :

- **Harmonisation des valeurs** : certaines colonnes comme `Sleep Duration` contenaient des guillemets superflus (ex : `'5-6 hours'`) qui ont été supprimés.

- **Typage correct des variables** : les colonnes catégorielles (ex. : `Gender`, `City`, `Profession`, `Degree`) ont été converties en facteurs. De même, les variables binaires (`Suicidal Thoughts`, `Family History of Mental Illness`, `Depression`) ont été recodées pour faciliter les analyses statistiques.
- **Renommage des colonnes** : certaines colonnes ont été renommées pour éviter les problèmes liés aux caractères spéciaux (comme les points ou espaces).
- **Contrôle des valeurs manquantes** : aucune valeur manquante n'a été détectée dans le jeu de données, ce qui a permis de conserver l'ensemble des 27 901 observations.
- **Préparation à l'analyse exploratoire** : les données sont désormais prêtes à être analysées en profondeur, notamment via des résumés statistiques, des visualisations, et des modèles prédictifs.

Ces étapes permettent de garantir la robustesse des analyses statistiques ultérieures et d'éviter les biais dus à des erreurs de codage ou à des valeurs aberrantes.

Scientific Workflow

L'approche adoptée dans cette étude suit une démarche rigoureuse de science des données en plusieurs étapes distinctes mais complémentaires. Le travail a débuté par une phase d'importation et d'exploration initiale du jeu de données, afin d'en comprendre la structure et d'identifier d'éventuelles incohérences, valeurs manquantes ou erreurs de saisie. Cette phase exploratoire a permis de guider les choix de nettoyage de données.

Ensuite, une procédure de nettoyage ciblée a été appliquée : transformation des types de variables (facteurs, numériques), uniformisation des modalités, détection et traitement des valeurs manquantes ou aberrantes. Chaque traitement a été réalisé de manière transparente dans le cadre de la programmation lettrée, afin d'assurer la traçabilité et la reproductibilité complète de l'analyse.

Une fois les données prêtes, des techniques de visualisation et d'analyse statistique ont été mobilisées pour explorer les relations potentielles entre les facteurs étudiés (académiques, sociaux, personnels) et la variable d'intérêt principale : la présence de symptômes dépressifs. Le raisonnement s'est appuyé sur une logique inductive, fondée sur l'observation des régularités dans les données, pour identifier les profils les plus vulnérables.

Enfin, les résultats ont été interprétés à la lumière du contexte psychosocial et universitaire, en tenant compte des limites possibles liées à la qualité ou la nature des données collectées.

Data Representation Choices

La visualisation des données a joué un rôle central dans cette étude pour faciliter la compréhension des tendances, différences et associations entre variables.

Plusieurs types de graphiques ont été utilisés, en fonction de la nature des variables à représenter :

- **Boxplots** : ils ont permis de comparer la distribution du CGPA selon le statut dépressif, en mettant en évidence les éventuelles différences de performance académique entre étudiants avec ou sans symptômes dépressifs.
- **Diagrammes en bâtons** : utilisés pour représenter le taux de dépression en fonction de différents facteurs catégoriels comme le genre, la ville, le stress financier ou les antécédents familiaux. Ces représentations facilitent la lecture directe des variations de prévalence.
- **Barres empilées** : elles ont permis d'analyser la répartition de la dépression selon certaines modalités spécifiques, comme les habitudes alimentaires, en visualisant la proportion d'étudiants concernés au sein de chaque catégorie.
- **Graphiques en courbes** : deux types de courbes ont été employés. D'une part, pour illustrer l'évolution du taux de dépression en fonction de facteurs groupés. D'autre part, pour observer les tendances différenciées par genre et domaine d'étude.

Ces choix visuels ont été motivés par la volonté de rendre l'analyse à la fois accessible, précise et interprétable, en maximisant la clarté des comparaisons.

Analysis in Literate Programming

L'objectif de cette section est d'identifier les facteurs les plus déterminants dans la survenue de la dépression chez les étudiants à partir des données disponibles, en adoptant une approche de programmation lettrée (*literate programming*). Cela implique de combiner analyses statistiques, visualisations graphiques et explications textuelles afin de construire un raisonnement rigoureux, traçable et reproductible.

Nous avons sélectionné un ensemble de variables susceptibles d'être liées à l'état psychologique des étudiants sur la base d'hypothèses théoriques et du contexte universitaire. Chaque facteur a été étudié séparément, à travers des visualisations adaptées, puis analysé afin d'évaluer son lien potentiel avec la variable cible : la présence ou non de symptômes dépressifs.

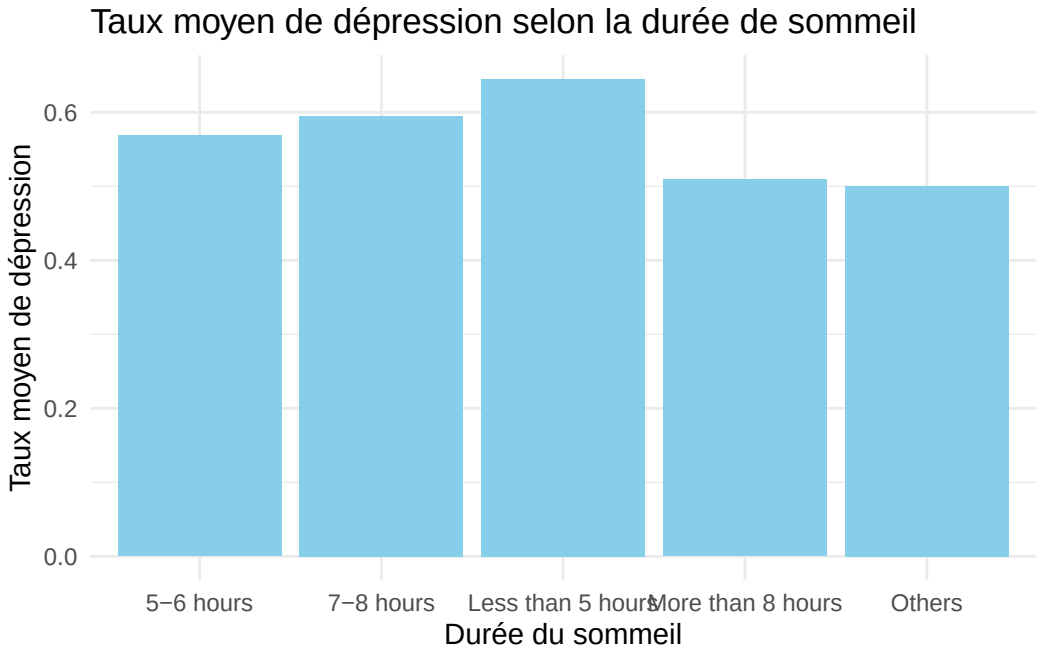
Les principaux facteurs étudiés sont les suivants :

- **La durée moyenne de sommeil**
- **Le niveau de pression académique**
- **Le niveau de satisfaction vis-à-vis des études**
- **Les habitudes alimentaires**
- **Le niveau de stress financier**
- **La performance académique (CGPA)**
- **L'âge, le genre, le domaine d'étude**

Pour chacun de ces facteurs, nous justifierons brièvement leur inclusion, visualiserons les données à l'aide de graphiques, interpréterons les résultats, puis conclurons sur leur importance relative dans la prédiction de la dépression.

1. La durée moyenne de sommeil

Le sommeil joue un rôle fondamental dans la régulation des émotions, la concentration et la santé mentale globale. Chez les étudiants, des rythmes de sommeil perturbés ou une privation chronique de sommeil peuvent contribuer à une vulnérabilité accrue à la dépression. Il est donc pertinent d'étudier si la durée moyenne de sommeil est associée à un taux plus ou moins élevé de dépression dans notre population.



Le graphique ci-dessus représente le taux moyen de dépression au sein de notre population étudiante en fonction de la durée de sommeil déclarée.

1. Tendance générale

On observe une relation inverse assez nette : **plus la durée de sommeil augmente, plus le taux moyen de dépression diminue**.

2. Points clés

   - $< 5\ h$ *de sommeil* : pic de prévalence de dépression (65 %)
   - *5–6 h* : léger creux à 56 %, avant une petite remontée

- *7–8 h* : taux d'environ 60 %, supérieur à la catégorie "5–6 h".
- *> 8 h* : chute marquée du taux (50 %), retrouvée également dans la catégorie "Others" (48 %).

3. Interprétation

Les très courtes durées de sommeil (< 5 h) sont associées à un risque de dépression nettement plus élevé. Un effet seuil apparaît : au-delà de 8 h, le taux de dépression se rapproche des niveaux les plus faibles, suggérant un effet protecteur du sommeil prolongé.

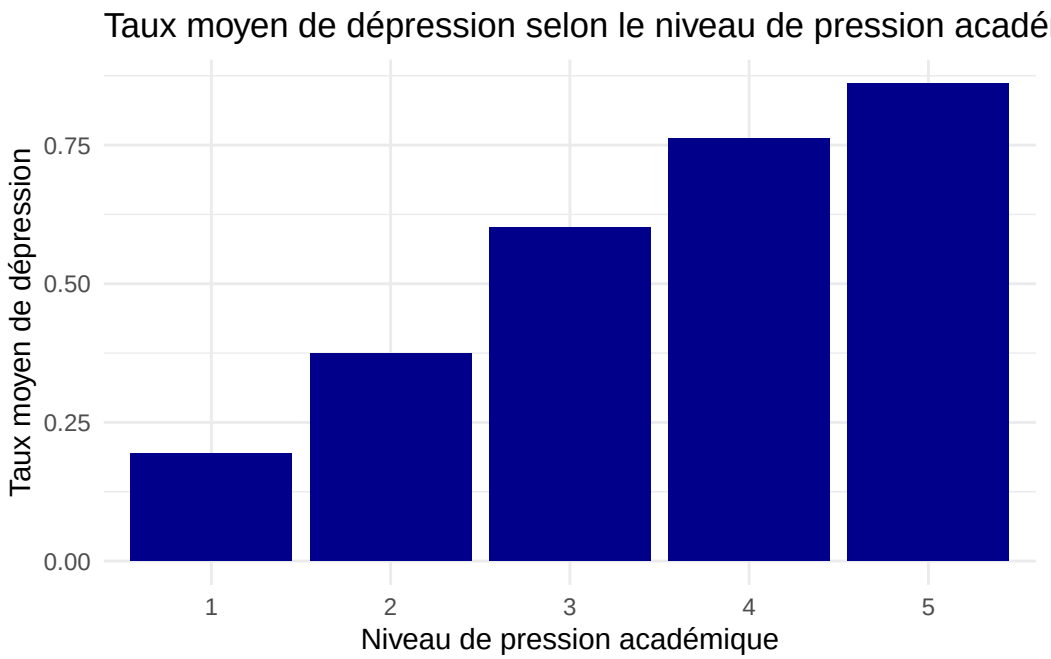
La légère hausse observée pour la tranche "7–8 h" par rapport à "5–6 h" pourrait traduire des effets de groupe ou de biais de reporting et mérite un examen plus fin (taille d'échantillon, variabilité individuelle).

4. Conclusion provisoire

Les données suggèrent qu'un sommeil insuffisant (< 5 h) constitue un facteur de risque majeur de dépression chez les étudiants, tandis qu'un temps de repos supérieur à 8 h apparaît comme potentiellement protecteur.

2. Le niveau de pression académique

La pression académique est un facteur déterminant dans la santé mentale des étudiants. En milieu universitaire, les exigences de performance, les échéances multiples et la compétition entre pairs peuvent entraîner une augmentation significative du stress psychologique. L'explosion prolongée du stress psychologique peut être corréler à des symptômes dépressifs. Dans le contexte de ce projet, analyser l'impact de la pression académique sur le taux de dépression permet de comprendre dans quelle mesure le système éducatif contribue à la détérioration du bien-être des étudiants, et peut suggérer des pistes pour des politiques de prévention ou de soutien psychologique dans les établissements d'enseignement supérieur.



Le graphique illustre la relation entre *l'intensité de la pression académique (de 1 = très faible à 5 = très forte) et le taux moyen de dépression* observé chez les étudiants.

1. Tendance claire et monotone

   - À niveau 1 (pression très faible), le taux moyen de dépression est extrêmement bas (20 %).
   - À niveau 2, on passe déjà à 38 %.
   - De niveau 3 à niveau 5, on observe une montée régulière :
     - Niveau 3 : 60 %
     - Niveau 4 : 76 %
     - Niveau 5 : 86 %

2. Points clés

La hausse du taux de dépression est non linéaire : l'écart est faible entre 1→2 (+18 pp), puis s'élargit à chaque palier (2→3 +22 pp, 3→4 +16 pp, 4→5 +10 pp). Le palier critique semble se situer dès le niveau 3 où la majorité des étudiants (6 sur 10) présentent des signes de dépression.

3. Interprétation

Pression académique modérée à élevée (3) apparaît comme un facteur de risque majeur, avec plus de 60 % de prévalence. Un stress léger (niveau 2) multiplie déjà par 2 le risque par rapport à un stress quasi nul (niveau 1). Le passage de 4 à 5 montre un effet de plateau, suggérant

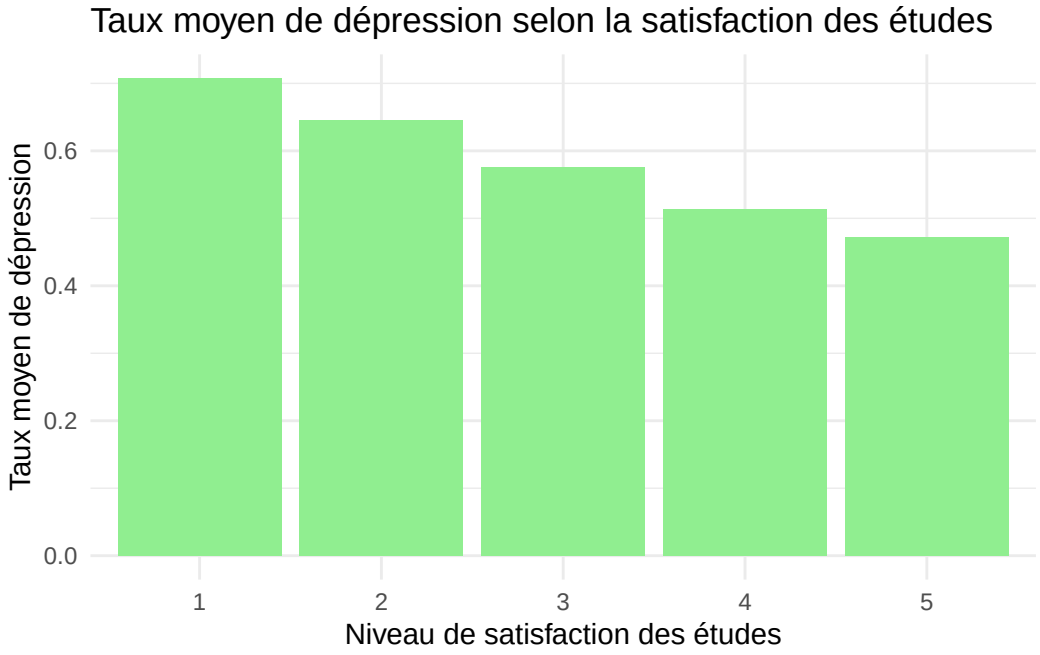
que, au-delà d'un certain niveau, l'augmentation de la pression aggrave encore la dépression, mais de façon un peu moins marquée qu'entre 2 et 3.

4. Conclusion provisoire

Le risque de dépression augmente fortement avec la pression académique : au-delà du niveau 3, plus de la moitié des étudiants sont touchés, indiquant la nécessité d'interventions ciblées pour réduire le stress lié aux études.

3. Le niveau de satisfaction vis-à-vis des études

La satisfaction vis-à-vis des études reflète l'adéquation entre les attentes de l'étudiant et sa réalité académique. Un faible niveau de satisfaction peut entraîner une démotivation, une perte de sens et un mal-être émotionnel. Ce facteur est donc pertinent pour évaluer son lien potentiel avec l'apparition de la dépression.



Le graphique représente la *prévalence moyenne de la dépression en fonction du niveau de satisfaction académique (échelle de 1 "très insatisfait" à 5 "très satisfait").*

1. Relation inverse claire

   - Satisfaction = 1 (très faible) : taux moyen de dépression le plus élevé, autour de 70 %.
   - Satisfaction = 2 : baisse à environ 64 %.

- Satisfaction = 3 : nouvelle diminution à 57 %.
- Satisfaction = 4 : taux intermédiaire, 51 %.
- Satisfaction = 5 (très élevée) : taux le plus bas, 48 %.

2. Points saillants

   - Le passage de 1→2 (−6 pp) est moins marqué que de 2→3 (−8 pp), suggérant que la transition du très faible au faible niveau de satisfaction a un impact modéré, mais que l'amélioration de 2 à 3 est associée à une baisse plus forte de la dépression.
   - Au-delà de 3, chaque point de satisfaction supplémentaire réduit encore le taux, mais de manière plus progressive (−6 pp puis −3 pp).

3. Interprétation

   - Un sentiment d'insatisfaction (2) correspond à une prévalence très élevée de dépression (> 60 %).
   - Un seuil critique semble autour de la note 3, point à partir duquel la prévalence passe sous la barre des 60 %.
   - La progression positive de la satisfaction (4–5) est protectrice, ramenant le taux sous 50 %.

4. Conclusion provisoire

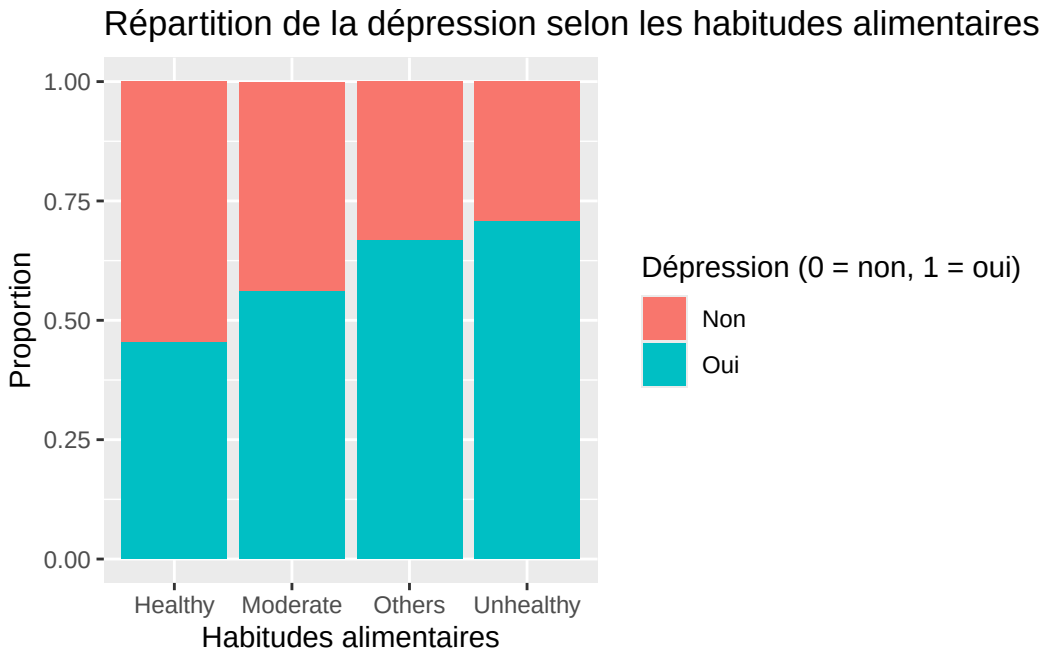
La diminution progressive du taux de dépression avec l'augmentation de la satisfaction des études indique un effet protecteur notable : *améliorer le bien-être et la satisfaction académique pourrait contribuer à réduire significativement la prévalence de la dépression chez les étudiants.*

4. Les habitudes alimentaires

Les *habitudes alimentaires* influencent directement la santé physique et mentale. Une alimentation déséquilibrée peut entraîner des carences (en vitamines, acides gras essentiels, etc.) qui augmentent les risques de troubles de l'humeur comme la dépression. Il est donc pertinent d'examiner leur lien avec la prévalence de la dépression.

- Pourquoi un graphique à barres empilées (proportionnelles) ?

Ce type de graphique permet de visualiser la proportion de cas de dépression (oui/non) au sein de chaque catégorie d'habitudes alimentaires. Cela rend la comparaison des pourcentages relative plus claire qu'un simple comptage.



Ce graphique à barres empilées compare pour chaque catégorie d'habitudes alimentaires ("Healthy", "Moderate", "Others", "Unhealthy") la proportion d'étudiants dépressifs (en turquoise) vs non dépressifs (en saumon).

1. Tendances principales

   - Healthy (alimentation saine) : Environ 45 % de dépressifs, 55 % de non-dépressifs.
   - Moderate (alimentation modérée) : ~ 55 % de dépressifs, 45 % de non-dépressifs.
   - Others (habitudes atypiques ou données manquantes) : ~ 65 % de dépressifs, 35 % de non-dépressifs.
   - Unhealthy (alimentation peu équilibrée) : 70+ % de dépressifs, 30-35 % de non-dépressifs.

2. Interprétation

   - Gradient clair : plus les habitudes alimentaires se détériorent, plus la proportion d'étudiants dépressifs augmente.
   - Effet protecteur potentiel d'une alimentation saine : la catégorie "Healthy" affiche le taux le plus bas de dépression.
   - Les étudiants avec des habitudes "Others" (potentiellement irrégulières ou non classées) présentent un taux de dépression très élevé, presque équivalent à "Unhealthy".

3. Limites

- Taille de chaque groupe non indiquée : les proportions peuvent être influencées par un faible effectif dans certaines catégories.
- Confondants non contrôlés : stress académique, sommeil, etc., pourraient aussi varier selon le régime alimentaire.
- Données auto-rapportées : biais possible dans la déclaration du régime alimentaire.

4. Perspectives d'approfondissement

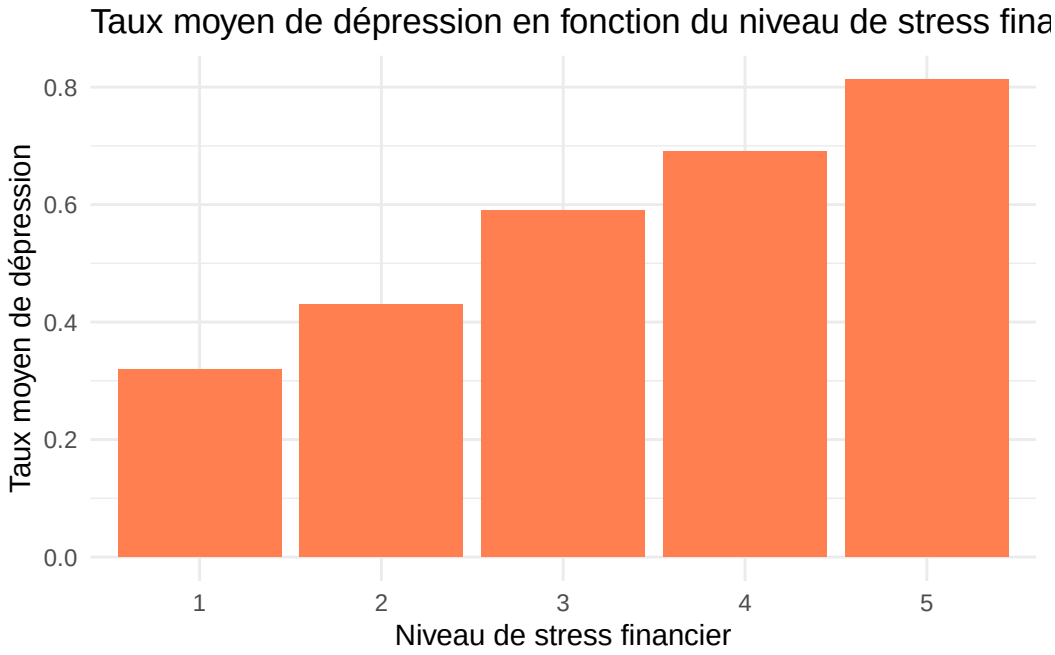
   - *Test statistique (^{2})* pour vérifier que la répartition diffère significativement selon les groupes alimentaires.
   - *Modèle multivarié (logistique)* contrôlant pour d'autres facteurs de risque.
   - *Analyse qualitative ou enquêtes complémentaires* pour comprendre pourquoi les étudiants à "Others" ont un risque si élevé.

5. Conclusion provisoire

Ce graphe suggère un lien fort entre qualité de l'alimentation et prévalence de la dépression chez les étudiants, avec un effet protecteur notable pour une diète saine et une augmentation marquée du risque pour des habitudes moins équilibrées.

5. Le niveau de stress financier

Le stress financier est un facteur connu pour affecter la santé mentale, notamment chez les étudiants, souvent confrontés à des difficultés économiques. Il peut amplifier l'anxiété, réduire la qualité de vie et nuire à la concentration académique. L'analyse de ce facteur permet d'évaluer dans quelle mesure les pressions économiques sont liées à la dépression étudiante.



Ce graphique présente *le taux moyen de dépression pour chaque niveau de stress financier (niveau de stress de 1 à 5).*

1. Tendance claire et monotone

   - Stress = 1 (très faible) : taux 32 % (le plus bas)
   - Stress = 2 : taux 43 %
   - Stress = 3 : taux 59 %
   - Stress = 4 : taux 69 %
   - Stress = 5 (très élevé) : taux 82 % (le plus haut)

On observe une corrélation fortement positive : **plus le stress financier augmente, plus la prévalence moyenne de la dépression grimpe**, passant de ~ 32 % pour un stress quasi nul à ~ 82 % au niveau maximal.

2. Points saillants

   - La transition 1→2 (+ 11 points) marque déjà une augmentation notable.
   - Le palier 2→3 voit un bond encore plus important (+ 16 points).
   - Au-delà de 3, chaque point de stress supplémentaire augmente le taux d'environ + 10 points en moyenne.

3. Interprétation

- Un stress financier modéré (3) est associé à une prévalence de dépression très élevée (> 50 %).
- Le seuil critique semble se situer autour de 3/5, au-delà duquel le risque devient majoritaire.
- Un stress financier maximal (5/5) fait basculer plus de 8 étudiants sur 10 vers la dépression.

4. Conclusion provisoire :

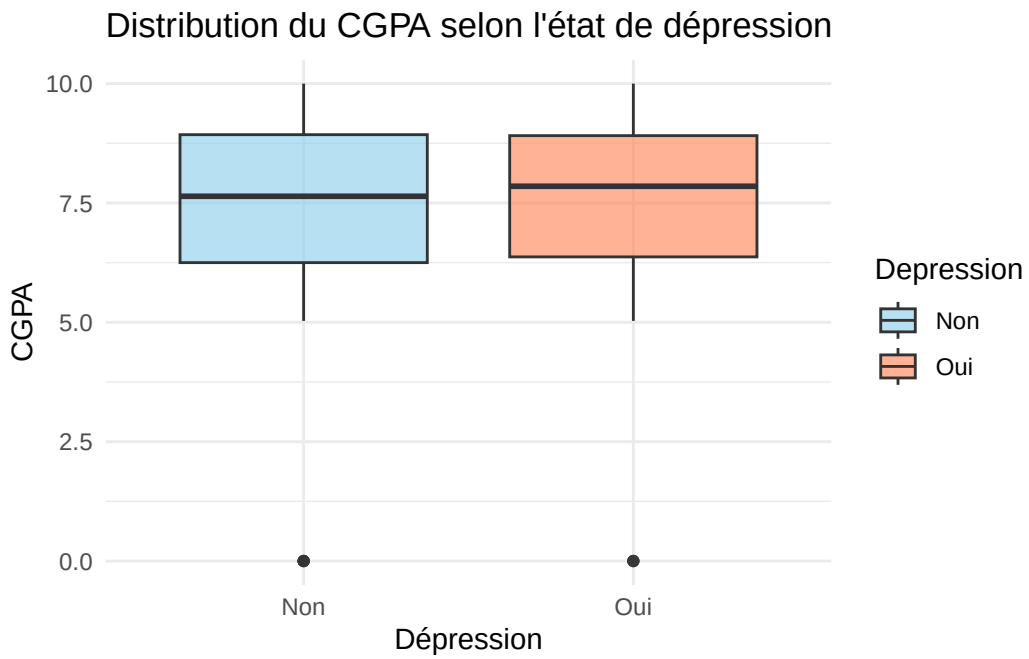
Le stress financier apparaît comme l'un des facteurs de risque les plus puissants pour la dépression chez les étudiants ; toute intervention visant à soulager ou accompagner la gestion du budget étudiant pourrait donc avoir un impact majeur sur leur santé mentale.

6. La performance académique (CGPA)

Le **CGPA (Cumulative Grade Point Average)** est un indicateur direct de performance académique. Des études suggèrent que de faibles performances peuvent augmenter le stress et les risques de détresse psychologique, y compris la dépression. Comparer les CGPA entre étudiants dépressifs et non-dépressifs peut révéler si une corrélation existe entre la réussite scolaire et l'état psychologique.

- Pourquoi un boxplot ?

Le boxplot est l'outil idéal pour comparer la distribution d'une variable continue (CGPA) entre deux groupes (dépression : oui/non). Il montre la médiane, les quartiles, les extrêmes et les valeurs aberrantes, ce qui permet d'identifier les tendances centrales et les dispersions éventuelles entre groupes.



Ce graphique compare la répartition des notes CGPA (sur une échelle 0–10) pour deux groupes :

- Non-dépressifs
- Dépressifs

1. Description générale

   - Chaque boîte (boxplot) montre le 1er quartile (Q1), la médiane, le 3 quartile (Q3) et les « moustaches » (valeurs extrêmes non-outliers).
   - Les points isolés (à CGPA = 0) sont des valeurs manquantes ou erronées.

2. Points saillants

   - Médianes très proches :
     - Non-dépressifs 7,6
     - Dépressifs 7,8
   - Étendues (Q1 → Q3) comparables :
     - Groupe non-dépressif : de 6,5 à 9,0
     - Groupe dépressif : de 7,0 à 8,8
   - Outliers à 0 pour les deux groupes, indiquant un ou deux enregistrements de CGPA invalides ou manquants.

3. Interprétation

    - Absence d'écart significatif entre les deux distributions : les étudiants dépressifs et non-dépressifs présentent des CGPA très similaires en termes de médiane et de dispersion.
    - Performance académique (CGPA) ne semble pas être, dans cette population, un facteur différenciateur majeur de la dépression.

4. Conclusion provisoire

La distribution du CGPA est très proche entre étudiants dépressifs et non-dépressifs, suggérant que, dans ce jeu de données, la performance académique n'est pas un indicateur principal de la dépression.