1. Introduction

Aujourd'hui, l'un des langages les plus utilisés (si ce n'est le plus utilisé) pour analyser des données est Python. Le but de ce mini-projet est de permettre de fournir une analyse détaillée à partir de données dit "en vrac" récupérées d'un dataset public. Ainsi, le langage Python va nous permettre de réaliser cette analyse, en récupérant des résultats à interpréter et donner un sens à une étude. Pour cela, dans ce présent rapport, je vais vous expliquer mes choix et les étapes passées ainsi que vous présenter les différents résultats que j'ai obtenus.

2. Choix du dataset

Le choix du dataset est très important. Non seulement il doit contenir assez de lignes de données pour être pertinent mais en plus il doit être signifiant pour moi, c'est-à-dire que ça doit concerner un sujet qui me plaît et qui me touche. Ainsi, l'analyse n'en sera que plus complète et surtout beaucoup plus intéressante que si je prenais n'importe quel dataset. Dans cette optique, j'ai donc pris un dataset qui reprenait les réponses à un questionnaire concernant "le stress des étudiants". C'est un sujet qui me concerne étant aussi étudiante. On peut aussi ajouter à cela que cette étude fournit deux datasets différents, soit beaucoup plus de choix et beaucoup plus de libertés sur les analyses à effectuer. Pour finir, après un premier coup d'œil, on peut voir que les données sont pour la plupart d'ordre numérique, et donc beaucoup plus simple à exploiter par rapport à d'autres. Ainsi, je me dirige vers cette étude afin de commencer mes analyses.

3. Étapes d'analyse

Première chose, il faut commencer par extraire nos données, les convertir et les nettoyer. Pour cela, on a accès à la librairie Pandas sous Python qui nous permet de convertir des fichiers de données (en l'occurrence, ici, en csv), en tableau de données. L'avantage de Pandas, c'est qu'il nous fournit énormément d'outils pour exploiter les tableaux de données qu'elle crée, et ces tableaux de données sont facilement malléables. Ainsi, il nous suffit d'utiliser la fonction "replace" ainsi que les propriétés de tableau Python pour clarifier nos données. D'abord on transforme les 0 et 1 de la colonne "Gender" en "male" pour les 0 et "female" pour les 1. Ensuite, on se concentre sur les étudiants entre 18 et 25 ans, car les données comportent aussi des personnes d'âges plus diversifiées mais très spécifiques, ce qui peut, je pense, fausser certains résultats. Enfin après tout ça, je supprime toutes les lignes qui possèdes des données manquantes, encore une fois pour éviter de fausser les résultats.
Après avoir nettoyé nos tableaux de données (car je le rappelle nous avons 2 datasets), désormais, on va commencer nos premières analyses très simples pour avoir une vue un peu plus complète de l'étude. On prend donc les mesures statistiques classiques, en prenant en compte que les valeurs données sont comprises entre 0 et 5 : la parité hommes/femmes, la moyenne d'âge des personnes interrogées, l'écart-type d'âge entre les personnes interrogées, la moyenne de niveau de stress ressenti entre toutes les personnes interrogées, la moyenne du niveau de problème de sommeil au global et ensuite par âge et la répartition du type de stress ressenti par les étudiants.
Cela nous donne une direction sur les visualisations que l'on souhaite créer afin de peaufiner notre analyse.
Ainsi nous pouvons partir sur créer nos visualisations, qui seront bien plus parlantes pour atteindre l'objectif de l'étude, c'est-à-dire comprendre le stress chez les étudiants.

Pour créer ces visualisations, je vais m'aider de 3 autres librairies Python : Matplot, seaborn et sklearn. Avec ces derniers, on va créer 8 visualisations différentes, dont les résultats vont nous intéresser juste après.

4. Résultats principaux

Grâce aux visualisations, on comprend qu'il y a une évolution dans le type de stress ressenti chez les étudiants au fil des années, de plus, le genre n'a quasiment aucun impact sur celui-ci. Avec la heatmap des corrélations entre les différentes caractéristiques du deuxième dataset, on comprend que le stress, l'anxiété, l'état dépressif ou encore les problèmes de sommeil sont souvent dûs à des facteurs externes, tels que les cours ou le harcèlement. De plus, on voit aussi, comme dit précédemment, que l'âge influe sur le type de stress ressenti. Plus, on grandit, plus le stress ressenti est "un bon stress" qui nous motive et nous pousse à faire mieux plutôt qu'une peur paralysante comme peuvent le ressentir les jeunes étudiants. Les étudiants plus jeunes encore sont plus insouciants et ressentent très peu de stress voire pas du tout. Le nuage de points de la dernière visualisation nous confirme que le stress ressenti par les personnes plus âgées est moins intense que pour les jeunes, la prise de maturité, le recul et l'expérience peuvent expliquer ce résultat.
Enfin la projection des individus nous indique que, selon le niveau de stress, les causes et les conséquences sont quasi similaires à quelques exceptions près.
Pour finir, les causes principales des étudiants sont diverses et variées et nécessitent une étude à encore plus grande échelle pour avoir des mesures pertinentes aussi en fonction de la culture, l'environnement et le pays résident.