POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

INF8008 - Prétraitement de données

Hiver 2025

Travail Pratique 2

Analyse descriptive des données

Durée	2 heures
Session	Hiver 2025
Lieu de réalisation	L-3714 - En ligne
Date de Remise	Samedi 15 Février 2025 à 23h59
Taille de l'équipe	**2 personnes**
Pondération	**9%**
Directives particulières	**1.** Tout retard dans la remise du compte-rendu entraîne automatiquement une pénalité comme discuté dans le plan de cours. **2.** Aucun retard de plus de 24 heures ne sera admis, la note de zéro sur neuf (0/9) sera attribuée aux étudiants concernés. **3.** Aucun compte-rendu ne sera corrigé, s'il est soumis par une équipe dont la taille est différente **de deux (2) étudiants** sans l'approbation préalable du chargé de laboratoire. La note de zéro sur neuf (0/9) sera attribuée aux étudiants concernés. **4.** Soumission du compte rendu par **Moodle** uniquement (https://moodle.polymtl.ca). **5.** Aucune soumission "hors **Moodle**" ne sera corrigée. La note de zéro sur neuf (0/9) sera attribuée aux étudiants concernés.

Table des matières

1. Travail à remettre

Vous devez remettre sur Moodle un fichier compressé .zip contenant :

1) Le code : Un Jupyter notebook en Python qui contient le code tel implanté avec les librairies minimales demandées pour ce TP (Python, Pandas, Matplotlib). Le code doit être exécutable sans erreur et accompagné des commentaires appropriés dans le notebook de manière. Tous vos résultats doivent être reproductibles avec le code dans le notebook. Attention, en aucun cas votre code ne doit avoir été copié de d'ailleurs.

2) Un fichier pdf représentant votre notebook complètement exécuté sous format pdf (obtenu via latex ou imprimé en pdf avec le navigateur). Assurez-vous que le PDF est entièrement lisible. Tutoriel youtube

ATTENTION: assurez-vous que votre fichier compressé .zip ne dépasse pas la taille limite acceptée sur Moodle.

2. Évaluation

Rubriques	**Points**
Implémentation correcte et efficace du code Réponses aux questions de réflexion ou d'analyse	19
Qualité du code (noms significatifs, structure, performance, gestion d'exception, etc.) et commentaires	1
Total de points	20

3. Environnement et outils nécessaires

Google Colab ou Notebook Jupyter.

4. Introduction

Le TD2 porte sur la transformation, la distribution et les statistiques descriptives. Nous survolons l'utilisation de fonctions de base de Pandas et de l'analyse de données numériques. Les données du fichier Alzheimer_s_Disease_and_Healthy_Aging_Data.csv sont des données publiques provenant d'enquêtes sur le vieillissement et la santé, faites par le Département de la Santé et des Services sociaux des États-Unis. Contrairement aux données du TP1 qui avaient été traitées au préalable, celles utilisées pour ce TP ne le sont pas. Vous devrez traiter les données brutes pour obtenir une version plus condensée, facilitant l'analyse des tendances et des sous-groupes de population.

Les champs du fichier de données « **Alzheimer_s_Disease_and_Healthy_Aging_Data.csv »** sont les suivants :

- **YearStart/YearEnd** : années de début et de fin des données
- **LocationAbbr** : abréviation du lieu
- **Class** : catégorie des données (ex. : Santé mentale)
- **Topic** : sujet spécifique (ex. : détresse mentale fréquente)
- **Question** : question étudiée
- **Data_Value_Unit**: unité de mesure des données (ex. : pourcentage)
- **Data_Value** : valeur des données collectées
- **StratificationCategory1 / Stratification1** : catégorie et détail de la première stratification (ex. : âge, genre)

- **StratificationCategory2 / Stratification2** : catégorie et détail de la deuxième stratification (ex. : race, ethnie)

Ces données servent de base pour explorer les tendances, identifier des corrélations, et mieux comprendre les facteurs liés aux maladies neurodégénératives et à la santé mentale des populations vieillissantes. Votre objectif dans ce TP sera de préparer ces données pour qu'elles soient prêtes pour une analyse approfondie.

Les librairies Python qui seront à utiliser pour ce TP sont pandas, numpy et matplotlib.

À noter qu'au niveau de chaque question, il est recommandé de copier le DataFrame obtenu à la question précédente dans un nouveau DataFrame.

5. Objectifs du laboratoire

Cette séance de laboratoire a pour but de permettre à l'étudiant(e) de :

- Se familiariser avec l'environnement Python
- Se familiariser avec les notebooks
- Se familiariser avec les librairies pandas et matplotlib
- Se familiariser avec l'analyse des données numériques avec pandas
- Se familiariser avec la visualisation de données avec matplotlib

6. Exercices

Complétez le fichier notebook « INF8008_TP2_H25.ipynb » fourni afin de répondre aux questions suivantes qui y sont demandées. Vous devrez soumettre vos réponses dans le notebook. Veuillez lire attentivement les consignes concernant les livrables ainsi que le code d'honneur.

1) Analyse des données numériques

A: Vous remarquerez que ce jeu de données est assez large, avec 284142 lignes et 31 colonnes. Avec des ensembles de données de cette taille, on peut souvent trouver des défauts, comme des doublons de lignes.
Vérifiez donc s'il existe des valeurs en double dans le DataFrame. **(2 points)**

B: Il est possible d'extraire la durée du sondage en soustrayant l'année de début de l'année de fin. Utilisez lambda, ainsi que cette soustraction, pour garder les lignes avec une durée de sondage de moins d'1 an. **(3 point)**

C: Maintenant que cette étape est faite, les colonnes YearStart et YearEnd contiennent la même information. Renommez une des deux colonnes à `Year`, et supprimez l'autre. **(2 points)**

D: Certaines colonnes contiennent des données redondantes ou inutiles pour notre analyse. Éliminez toutes les colonnes inutiles en ne conservant que celles mentionnées dans l'introduction. Combien de colonnes reste-t-il ? **(point)**

E: Comme vu dans le module 1, le prétraitement des données consiste à gérer les défauts des données collectées, comme les valeurs nulles. La colonne `*Data_Value*` est importante pour notre analyse.

Vérifiez donc s'il existe des données manquantes dans la colonne `*Data_Value*`. Quel est le pourcentage de valeurs manquantes ? **(3 points)**

F: Deux façons de traiter les données manquantes: les remplacer par la valeur médiane ou les éliminer complètement.

Il n'existe pas de solution unique ou meilleure. Tout dépend de l'analyse effectuée. Il est essentiel d'examiner les effets de chacun de ces choix sur l'analyse ultérieure. C'est pourquoi, dans ce TP, nous essayerons les deux méthodes.
Vous devez donc:

1. Créez deux copies de l'ensemble de données.

2. Supprimez les valeurs manquantes d'une des copies.

3. Remplacez les valeurs manquantes d'une autre copie par la médiane.

Affichez les nouveaux dataframes. Vous devriez avoir autour de 186595 lignes pour l'un et 274881 lignes pour l'autre. **(4 points)**

G: Plusieurs classes existent. On va évaluer la santé mentale "*Mental Health*". Filtrez les données de la colonne "class" pour la valeur "Mental Health", puis déterminez la moyenne de `Data_Value` par `Year` et `Topic`. **(2points)**

2) Visualisation de données

H: Il est temps de comparer la suppression des données manquantes vs leur remplacement par la médiane. Pour cela, affichez les valeurs moyennes de `Data_Value` par année, pour chaque groupe et chaque topic. **(3 points)**

Bon travail!