POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

INF8008 - Prétraitement de données

Hiver 2025

Travail Pratique 1

Analyse descriptive des données

Durée	2 heures
Session	Hiver 2025
Lieu de réalisation	L-3714 - En ligne
Date de Remise	Samedi 1 Février 2025 à 23h59
Taille de l'équipe	**2 personnes**
Pondération	**6%**
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 six (0/6) 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 six (0/6) 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 six (0/6) 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 TD1 porte sur l'analyse descriptive de données. Nous survolons l'utilisation de fonctions de base de Pandas et de l'analyse de données numériques et de leur visualisation. Les données du fichier **ntsb-accidents.csv** sont des données sur les accidents aériens que nous utiliserons pour les TP.

Les champs du fichier de données **ntsb-accidents.csv** sont les suivants :

- **event_id :** Identifiant unique de l'événement (souvent un code de référence pour suivre chaque cas).
- **ntsb_make :** Fabricant de l'appareil impliqué dans l'événement (exemple : BELL, ROBINSON).
- **ntsb_model :** Modèle spécifique de l'appareil (exemple : R22 BETA, R44).
- **ntsb_number :** Code d'identification de l'incident assigné par la NTSB.
- **year :** Année où l'événement a eu lieu.
- **date :** Date et heure exactes de l'événement (année/mois/jour heure:minute:seconde).
- **city :** Ville où l'événement s'est produit.
- **state :** État ou province où l'événement s'est produit.
- **country :** Pays où l'événement a eu lieu.
- **total_fatalities :** Nombre total de décès associés à l'incident.
- **latimes_make :** Fabricant simplifié utilisé dans les rapports (exemple : BELL, ROBINSON).
- **latimes_model :** Modèle simplifié utilisé dans les rapports (exemple : R22, 369).
- **latimes_make_and_model :** Combinaison du fabricant et du modèle pour décrire l'appareil de manière concise (exemple : BELL 407, ROBINSON R44).

Les librairies python qui seront à utiliser pour ce TP sont pandas et matplotlib.

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_TP1_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

Q1 : À l'aide de Pandas, chargez les données dans une variable nommée « df ». Quelle est la dimension de « df »? Combien y a-t-il de lignes et de colonnes? **(2 points)**

Q2 : Quelle est l'intervalle d'années et l'ensemble des fabricants et modèles d'hélicoptères compris dans le jeu de données? **(1 point)**

Q3 : Combien de décès y a-t-il en moyenne? **(1 points)**

Q4 : Combien d'**accidents** y a-t-il eu pour chaque combinaison de fabricant et modèle d'hélicoptère? Créez un tableau de fréquence du nombre d'accidents par combinaison de fabriquant et modèle. **(3 point)**

Q5 : Déterminez la moyenne, la médiane et l'écart-type du nombre de décès. Dans un premier temps, fournissez les statistiques avec AIRBUS 350. Puis, dans un second temps, fournissez les statistiques avec ROBINSON R44. **(3 points)**

Q6 : Combien d'accidents ont enregistré plus d'un décès? Combien de décès se sont produits en Californie en 2015?**(1 point)**

2) Visualisation de données

Q7 : Affichez le nombre de décès par année. Nommez vos axes et donnez un titre à votre graphique. Qu'observez-vous? Quelle année marque le plus de décès? Quelle année marque le moins? **(3 points)**

Q8 :

En utilisant la fonction subplot de matplotlib:

1. Affichez le nombre d'accidents par année dans un diagramme à barres.

2. Affichez le nombre de décès par année dans un diagramme à barres.

Nommez vos axes et donnez un titre à votre graphique.

En comparant ceux-ci, observez-vous une relation entre le nombre de décès et d'accidents? Pouvez-vous justifier la cause du nombre de décès minimal de l'année évoquée dans la question 7? **(5 points)**

Bon travail!