Projet de cours : Analyse exploratoire et qualité des données

Objectif général

Ce projet vise à vous faire mettre en pratique les notions vues en cours à travers l'analyse exploratoire d'un jeu de données réel ou réaliste. Vous utiliserez pour cela les outils Python et les bibliothèques apprises (pandas, matplotlib, seaborn, SciPy, etc.).

Vous aurez également à produire une analyse de la qualité des données, en appliquant différentes mesures de détection et de traitement des défauts les plus courants (valeurs manquantes, aberrantes, doublons, etc.).

Sujet et choix du jeu de données

Vous êtes libres de choisir un sujet qui vous intéresse : éducation, santé, sport, climat, transport, économie, société, etc.

Critères pour le jeu de données :

- Doit contenir au moins 10 colonnes.

- Doit inclure à la fois des variables numériques et catégorielles.

- La taille du jeu de données doit être suffisante pour permettre des analyses significatives (idéalement au moins 500 lignes).

Quelques sources de données ouvertes recommandées:

- Open Data Canada: https://open.canada.ca/en/open-data

- Données Québec: https://www.donneesquebec.ca/fr/

- Kaggle Datasets: https://www.kaggle.com/datasets

Contenu attendu du livrable

1. Présentation du projet

- Contexte du sujet choisi
- Objectifs de l'analyse (formuler une ou plusieurs hypothèses à tester)
- Présentation du jeu de données (source, description, structure)

2. Analyse exploratoire

- Statistiques descriptives : moyenne, médiane, écart-type, fréquence des catégories, etc.
- Visualisation des distributions et des corrélations :
- - Histogrammes, boxplots, nuages de points, heatmaps, etc.
- - Minimum cinq graphiques justifiés (univariés et bivariés)

3. Nettoyage et prétraitement des données

- Suppression des doublons
- Gestion des valeurs manquantes : suppression ou imputation
- Traitement des valeurs aberrantes : détection et décision (suppression/imputation)
- Encodage des variables catégorielles : Label Encoding et/ou One-Hot Encoding
- Normalisation ou standardisation des variables numériques : MinMaxScaler, StandardScaler, etc.

4. Analyse statistique

- Vérification de vos hypothèses à l'aide de :
- - Tests statistiques (ex. test t, test du chi², etc.)
- - Corrélations (Pearson, Spearman…)
- Interprétation et justification des résultats

Analyse de la qualité des données

Vous devez faire l'analyse des données et produire une document dédiée à l'évaluation de la qualité des données utilisées, à travers les critères suivants :

Critère	Définition	Évaluation attendue
Exactitude	Les données reflètent-	Exemples d'erreurs

	elles la réalité observée ?	détectées, corrections apportées
Complétude	Y a-t-il des valeurs manquantes critiques ?	Mesures de taux de complétude, impact sur l'analyse
Cohérence	Les données sont-elles homogènes dans leur format et contenu ?	Incohérences détectées (ex : formats mélangés, doublons)
Représentativité	L'échantillon est-il représentatif de la population cible ?	Biais potentiels, couverture des cas particuliers

👥 Travail en équipe

Le projet peut être réalisé seul ou en binôme. Si vous travaillez à deux, chaque membre doit pouvoir contribuer significativement au travaux.

📝 Livrables

- Un fichier Jupyter Notebook (`projet_analyse_donnees_nom(s).ipynb`)

- Document de l'analyse de qualité de données

🧮 Échéance

🗓️ Date de remise: Fin du cours