



Agence Nationale de la Statistique et de la Démographie
(ANSD)

Ecole nationale de la Statistique et de l'Analyse
économique - Pierre Ndiaye (ENSAE)

ISEP3

Projet Python Avancé : Analyse statistique descriptive et inférentielle d'un jeu de données

Objectif :

Réaliser une analyse statistique complète sur un jeu de données réel ou fictif, en utilisant Python et les bibliothèques **Pandas**, **Numpy** et **Matplotlib**. Effectuer également des tests statistiques pour répondre à des hypothèses prédéfinies.

Étapes du projet :

1. Chargement et exploration des données :

- Importer le jeu de données (au format .csv, .xlsx, etc.).
- Afficher les 5 premières lignes, les types de données, et vérifier les valeurs manquantes.
- Résumer les statistiques descriptives pour chaque colonne numérique (moyenne, médiane, écart-type, etc.).

Exemple de données possibles :

- Données sur les performances académiques des étudiants.
- Données économiques (PIB, taux de chômage, etc.).
- Données climatiques (température, précipitations, etc.).

2. Visualisation des données :

- Histogrammes pour visualiser la distribution des variables.
- Boxplots pour détecter les outliers.
- Scatter plots pour observer les relations entre deux variables.

3. Statistique descriptive avancée :

- Calculer les mesures de tendance centrale et de dispersion pour certaines colonnes.
- Analyser la corrélation entre les variables numériques et créer une matrice de corrélation avec un heatmap.

4. Tests statistiques :

- Effectuer un test de normalité sur une variable (exemple : test de Shapiro-Wilk ou test de Kolmogorov-Smirnov).
- Comparer deux groupes avec un test t de Student ou un test de Mann-Whitney.
- Tester l'indépendance entre deux variables catégorielles avec un test du Chi-carré.
- Calculer et interpréter un intervalle de confiance pour une moyenne.

5. Rapport et présentation :

- Rédiger un rapport contenant :
 - Les résultats des analyses descriptives.
 - Les conclusions des tests statistiques.
 - Les visualisations générées.
- Présenter leurs conclusions sous forme d'une courte présentation (facultatif).

Exigences :

- Utiliser les bibliothèques suivantes :
 - pandas pour la manipulation des données.
 - Numpy pour les calculs statistiques.
 - Matplotlib pour la visualisation.
 - (Optionnel) Seaborn pour des graphiques avancés.
- Chaque groupe doit utiliser un jeu de données différent de ceux d'un autre groupe.

Exemple de jeu de données :

Un fichier .csv avec les colonnes suivantes :

- Nom
- Âge
- Sexe (M/F)
- Note en Mathématiques
- Note en Science
- Temps d'étude par semaine (heures)
- Participation aux activités extrascolaires (Oui/Non)