Introduction à l'analyse de données avec Pandas

Dany Anderson & Ariste Yougbaré
Pole Data-IA
February 6, 2026

Contents

L'analyse de données

L'analyse des données est le processus d'inspection, de nettoyage, de transformation et de modélisation des données dans le but de découvrir des informations utiles, d'éclairer les conclusions et de soutenir la prise de décision .

04 étapes clés

- Inspection
- Nettoyage
- Transformation
- Modélisation

L'objectif etant d'assiter dans la prise de décision

Outils d'analyse de données

Rappel Rapide Concepts Statistiques

1. Moyenne (Mean)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Mesure la tendance centrale des données. Utilisée en EDA pour résumer une variable numérique.

2. Médiane (Median) Valeur centrale d'un ensemble trié. Robuste aux valeurs extrêmes, utile pour détecter une asymétrie.

3. Variance (Variance)

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Mesure la dispersion des données autour de la moyenne.

Rappel Rapide Concepts Statistiques

4. Écart-type (Standard Deviation)

$$\sigma = \sqrt{\sigma^2}$$

Dispersion exprimée dans les mêmes unités que les données. Pratique pour comparer la variabilité.

5. Covariance

$$\mathrm{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Indique si deux variables évoluent ensemble (positif ou négatif).

6. Corrélation (Pearson)

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Normalise la covariance entre -1 et 1. Utilisée en EDA pour explorer des relations linéaires.

Rappel Rapide Concepts Statistiques

7. Minimum / Maximum Bornes inférieure et supérieure des données. En EDA, utile pour détecter des valeurs extrêmes.

8. Étendue (Range)

$$\text{Range} = \max(x) - \min(x)$$

Mesure simple de dispersion. Complète la variance et l'écart-type.

9. Quantiles / Quartiles Divisent les données en parties égales. Exemple : Q1 (25%), Q2 = médiane (50%), Q3 (75%). Très utilisés pour résumer la distribution.

Rappel Rapide Concepts Statistiques

10. IQR (Interquartile Range)

$$\text{IQR} = Q3 - Q1$$

Mesure la dispersion centrale. En EDA, utilisé pour détecter les outliers (valeurs aberrantes).

Contents

Introduction à l'analyse de données avec python

Pourquoi Python ?

- Simple et intuitif
- Equipé de bibliothèques puissantes (Pas qu'en analyse de données)
- Outil gratuit et open source
- Une très large communauté autour du langage et de ses differents packages

Les bibliotheques python

- **Pandas** : Essentiel pour l'analyse exploratoire des donnnes
- **Matplotlib**: Utilisee pour la visualization des donnee
- **Numpy**: La bibliotheque de calculs numeriques de python
- **Seaborn**: Pour creer des visualizations avancees, concus en tant que surcouche de matplotlib
- **Scipy**: Calculs scientifiques complexes, incluant de l'optimization de fonctions, de l'algebre lineaire et du traitement d'image
- **Sci-kit Learn**: bibliotheque tres populaire pour faire du machine learning
- **Statsmodel**: Inclut des fonctions statistiques avancées

Introduction à Pandas

Pandas est une bibliothèque Python open-source conçue pour la manipulation et l'analyse de données

- Elle est largement utilisée en science des données et en analytique
- Elle est construite sur NumPy et fournit des structures de données et des outils efficaces pour travailler avec des données structurées.

Series et DataFrames

Series	**DataFrame**
Structure unidimensionnelle (1D)	Structure bidimensionnelle (2D)
Contient des données homogènes (même type)	Contient des données hétérogènes (types différents par colonne)
Indexée (chaque valeur possède une étiquette)	Indexée (chaque ligne possède une étiquette)
Peut avoir un nom optionnel	Colonnes nommées de façon obligatoire
Objet léger et simplifié (souvent une seule variable)	Objet riche et flexible (plusieurs variables en colonnes)

Contents

Practice time

Series

En Python (Pandas), une **Series** est une structure de données **unidimensionnelle** similaire à un tableau NumPy, mais enrichie d'un **index** permettant un accès plus flexible.

- **Caractéristiques :**
  - Contient une seule colonne de données
  - Chaque valeur est associée à un **index** (numérique ou personnalisé)
  - Accepte différents types de données : numériques, chaînes, booléens...
  - Supporte les opérations vectorisées comme NumPy

Notebook

- Analyse G7

DataFrame

Un **DataFrame** est une structure de données **bidimensionnelle** composée de lignes et de colonnes, comparable à une table ou une feuille Excel. C'est l'objet central de Pandas.

- **Caractéristiques :**
  - Organisation en lignes et colonnes
  - Chaque colonne est une Series
  - Index des lignes personnalisable
  - Robustesse pour la manipulation, le nettoyage et l'analyse de données

Notebook

- Analyse G7

Lecture CSV et gestion des valeurs manquantes

Objectifs :

- Charger un fichier CSV dans un DataFrame Pandas
- Identifier et traiter les valeurs manquantes

Méthodes clés :

- `pd.read_csv()` : lecture d'un fichier CSV
- `df.isna()` / `df.isnull()` : détection des valeurs manquantes
- `df.dropna()` : suppression des lignes/colonnes contenant des NA
- `df.fillna()` : remplacement des valeurs manquantes

Notebook

- Data
- Analyse Crypto
- Valeurs manquantes

Exemple d'analyse

Nous terminons avec l'analyse d'un petit jeu de données portant sur les salaires selon les filières universitaires.

Objectif : répondre, à partir des données, aux questions suivantes :

- Quelles filières offrent les salaires de départ les plus élevés ?
- Quelles filières mènent aux revenus les plus faibles après l'université ?
- Quelles filières présentent le plus fort potentiel de gains ?
- Quelles filières sont les moins risquées (faible variabilité des salaires) ?
- Entre Commerce, STIM et HASS, quel domaine rapporte en moyenne le plus ?

Data

- Toutes les donnnées externes sont disponibles via ce lien

Conclusion

Un peu plus...

- Visualisations avançées

Merci pour votre aimable attention