



Semaine 2 – Lire et interroger des données avec pandas

Outils informatiques pour les géosciences II

Sébastien Biass 

sebastien.biasse@unige.ch

1st floor, office 103

16 octobre 2026



Pandas & données tabulaires



Données tabulaires

NAME	COUNTRY	DATE	VAL
STROMBOLI	ITALY	2025-09-25	1
ONTAKE	JAPAN	2014-09-27	3
ST HELENS	USA	1980-05-18	5
PINATUBO	PHILIPPINES	1991-06-15	6

Structure

- **Ligne**: Observation
- **Colonnes**: Propriétés de l'observation

Format

- Excel, csv, base de données, shapefiles

Pandas

- **pandas** est une bibliothèque Python conçue pour **les données organisées en tableaux**.
- Elle permet de transformer une **question** en une **suite d'opérations explicites**:
 - comprendre la structure des données
 - choisir les observations et variables pertinentes
 - produire un résultat reproductible
 - conserver la trace des choix effectués



Raisonnement scientifique

La capacité de **pandas** à traiter les données **épaule toutes les étapes du raisonnement scientifique**:

1. **Question**: quel phénomène voulons-nous décrire?
2. **Données**: que représente une ligne et quelles variables sont disponibles?
3. **Sélection**: quelles observations répondent réellement à la question?
4. **Résultat**: que montre le tableau obtenu?
5. **Discussion**: quelle conclusion reste compatible avec les limites du catalogue?



Cinq actions sur les **données**

- **Inspecter**: Que contient le fichier? → *types de variables, valeurs manquantes*
- **Sélectionner**: Quelles variables sont nécessaires?
- **Filtrer**: Quelles observations appartiennent au groupe étudié?
- **Trier**: Quelles observations occupent les positions extrêmes?
- **Transformer**: Les variables utilisent-elles une représentation adaptée?

Anatomie d'un DataFrame

The diagram illustrates the structure of a DataFrame. A table with 4 columns and 4 rows is shown. The columns are labeled 'NAME', 'COUNTRY', 'DATE', and 'VEI'. The rows are labeled 'Row 0', 'Row 1', 'Row 2', and 'Row 3'. A green box labeled 'COLUMNS' has arrows pointing to each of the four columns. An orange box labeled 'ROWS' has arrows pointing to each of the four rows. Below the table, the columns are labeled 'COL 0', 'COL 1', 'COL 2', and 'COL 3'. To the right of the table, the rows are labeled 'Row 0', 'Row 1', 'Row 2', and 'Row 3'.

NAME	COUNTRY	DATE	VEI
STROMBOLI	ITALY	2025-09-25	1
ONTAKE	JAPAN	2014-09-27	3
ST HELENS	USA	1980-05-18	5
PINATUBO	PHILIPPINES	1991-06-15	6

- **Ligne**: une observation, une mesure, un échantillon ou un événement;
- **Colonne**: une variable/propriété décrivant les observations;



En Python, l'indexation commence à 0

Anatomie d'un DataFrame

INDEX = LABELS ALONG ROWS → df.index	NAME	COUNTRY	DATE	VEI	COLUMNS = LABELS ALONG COLUMNS → df.columns
	STROMBOLI	ITALY	2025-09-25	1	
	ONTAKE	JAPAN	2014-09-27	3	
	ST HELENS	USA	1980-05-18	5	
	PINATUBO	PHILIPPINES	1991-06-15	6	
		COL 0	COL 1	COL 2	

Row 0
Row 1
Row 2
Row 3

- **Ligne**: une *observation*, une *mesure*, un *échantillon* ou un *événement*;
- **Colonne**: une *variable*/*propriété* décrivant les observations;
- **Index** contient les *labels* des lignes;
- **Nom des colonnes** sont les *labels* des variables.

Principaux types de données

Type dans pandas	Exemple	Nature
<code>int64</code>	année, nombre d'événements	quantitative discrète ou catégorielle
<code>float64</code>	latitude, altitude	quantitative continue
<code>bool</code>	actif: vrai ou faux	catégorielle binaire
<code>string</code> / <code>object</code>	contexte tectonique, nom du volcan	catégorielle ou texte libre
<code>datetime64</code>	date d'une éruption	temporelle et ordonnée
<code>geometry</code> (GeoPandas)	point, ligne ou polygone	spatiale: position ou forme

- Variable **quantitative**: mesure une quantité sur laquelle les opérations numériques ont un sens
- Variable **catégorique**: distingue des groupes



Les données et les questions





Le catalogue du Global Volcanism Program

Le *Global Volcanism Program* du Smithsonian Institution rassemble des informations sur les volcans et leurs éruptions à l'échelle mondiale (Simkin et al. 2010).

Le fichier du cours contient:

- **678 volcans**
- **13 variables descriptives**
 - latitude, longitude, morphologie, altitude, tectonique, etc...
- uniquement les volcans dont l'année numérique de la dernière éruption connue est supérieure ou égale à 0





Quelques lignes du fichier

Name	Country	Morphology	Latitude	Longitude	Last Eruption	Elevation (m)
Campi Flegrei	Italy	Caldera	40.827	14.139	1538	458
Kilauea	USA	Shield	19.421	-155.287	2026	1222
Villarrica	Chile	Strato	-39.420	-71.930	2025	2847



Quelques lignes du fichier

Name	Country	Morphology	Latitude	Longitude	Last Eruption	Elevation (m)
Campi Flegrei	Italy	Caldera	40.827	14.139	1538	458
Kilauea	USA	Shield	19.421	-155.287	2026	1222
Villarrica	Chile	Strato	-39.420	-71.930	2025	2847

Localisation et activité

- | *Quels volcans indonésiens ont une éruption depuis 1900?*
- | *Combien de volcans du catalogue se trouvent dans l'hémisphère sud?*

Morphologie et altitude

- | *Quels volcans de l'hémisphère sud ont les sommets les plus élevés?*
- | *Comment la morphologie varie-t-elle avec le contexte tectonique?*



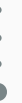
Une question, plusieurs opérations

Question:

| *Quels volcans de l'hémisphère sud ont les sommets les plus élevés?*

Raisonnement:

Décomposer en opérations simples avant de coder:





Une question, plusieurs opérations

Question:

| *Quels volcans de l'hémisphère sud ont les sommets les plus élevés?*

Raisonnement:

Décomposer en opérations simples avant de coder:

1. **Filtrer** les latitudes négatives
2. **Trier** l'altitude par ordre décroissant
3. **Sélectionner** le nom, le pays, l'altitude et l'année
4. **Inspecter** les premières lignes du résultat
5. **Discuter** ce que ce classement mesure réellement





Aujourd'hui



A vous

1. Créez un notebook `semaine_02.ipynb` dans le dossier `outils_info`
2. Téléchargez les données et placez-les dans le dossier `outils_info/data`
3. Suivre [Semaine 2](#) depuis le site
 - Copier/coller le code fourni (tout en expérimentant!)
 - Répondez aux [questions](#)
4. Utilisez les [cheat sheets](#) et la [documentation simplifiée](#) pour expérimenter et découvrir de nouvelles fonctionnalités
5. Répondez à la [Mini investigation](#) et documentez votre code avec Markdown

Question fil rouge

Comment réaliser une première exploration de données et les interroger dans un contexte de question de recherche?

À la fin de la séance, le notebook devra contenir:

- une question précise
- le code nécessaire pour isoler les données pertinentes
- un résultat lisible
- une observation fondée sur ce résultat
- une limite du catalogue ou de l'interprétation

Objectifs

À la fin de l'exercice, vous devriez pouvoir:

- ✓ Expliquer ce que représentent les lignes et les colonnes d'un tableau
- ✓ Vérifier la structure d'un fichier avant l'analyse
- ✓ Sélectionner, filtrer et trier les données nécessaires
- ✓ Relier chaque opération à une question précise
- ✓ Distinguer le résultat obtenu de son interprétation scientifique



References

Simkin T, Siebert L, Simkin T, Kimberly P (2010) Volcanoes of the World. University of California Press, Tucson, AZ

