Sustainability Analysis in Python

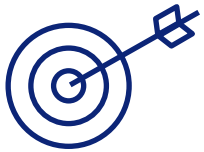
Pandas

Laura Scherer

Universiteit Leiden
The Netherlands

CML
Institute of Environmental Sciences

Learning Goals



- **Process unclean data**, describe datasets with metadata, and apply fair data principles
- Validate and assess uncertainties of models
- Test hypotheses and verify the underlying assumptions
- **Develop** clear and efficient **code in Python**, integrate user interaction, and keep track of versions

Cheat Sheet

Outline

Introduction

Data creation, input, and output

Data exploration

Summary statistics

Subset

Sorting

Data manipulations

Exercise

Introduction

Basic data types

1) Strings
2) Booleans
3) Integers
4) Floats

More complex data types

1) Lists
2) Dictionaries
3) NumPy arrays
4) Pandas Series and DataFrames

Introduction

- **Pandas**: fast and flexible tool for data analysis
- **GeoPandas**: extending pandas to allow for spatial operations with vector data / shapefiles (e.g. polygons like countries)

```
conda install geopandas
```

- Import of libraries

```
import pandas as pd
import geopandas as gpd
```

Outline

Introduction

Data creation, input, and output

Data exploration

Summary statistics

Subset

Sorting

Data manipulations

Exercise

Data creation, input, and output

Creating pandas data objects

```
series = pd.Series([1, 2, 3],
index = ['a', 'b', 'c'])


df = pd.DataFrame([[1, 2], [3, 4], [5, 6]],
index = ['i1', 'i2', 'i3'],
columns = ['col1', 'col2'])
```

Data creation, input, and output

Reading files

```
df = pd.read_csv('file_name.csv')
df = pd.read_excel('file_name.xlsx')
gdf = gpd.read_file('file_name.shp')
```

Many function arguments (see, e.g., here)

- e.g., `usecols = [0, 3, 9]`
- e.g., `skiprows = 3`
- e.g., `na_values = ['NA', 'null']`
- e.g., `dtype = {'col1':int, 'col2':float}`

Data creation, input, and output

Writing files

```
df.to_csv('file_name.csv')
```

Writing modes

- `mode = 'w'` → write
- `mode = 'a'` → append to existing file

Outline

Introduction
Data creation, input, and output
Data exploration
Summary statistics
Subset
Sorting
Data manipulations

Exercise

Data exploration

- `df.index # get row names`
- `df.columns # get column names`
- `df.shape # get number of rows and columns`
- `df.dtypes # return the data type (of each column)`
- `df.head() # return the first 5 rows`

```
pd.set_option('display.max_columns', None) # show all columns

pd.reset_option('display.max_columns') # return to default
```

Data exploration

- `df['column'].unique() # get unique values`
- `df.nlargest(5, 'column') # show the largest values`

- `pd.notna(df).all() # no missing values?`
- `pd.isna(df).sum() # count of missing values`
- `df1['column'].equals(df2['column']) # check equality`

Outline

Introduction
Data creation, input, and output
Data exploration
Summary statistics
Subset
Sorting
Data manipulations

Exercise

Summary statistics

- `df['column'].median()`
- `df['column'].mean()`
- `df.describe()`

  ```
  # summary statistics of numeric columns
  # count, number of unique values, most frequent string, and its frequency for string columns
  ```

- `df.groupby(df['column']).sum()`

Outline

Introduction
Data creation, input, and output
Data exploration
Summary statistics
Subset
Sorting
Data manipulations

Exercise

Subset

Selecting columns

- `df['column'] or df.column`
- `df[['column1', 'column2']]`

Selecting rows

- `df[df['column'] != n]`
- `df[(df['col1'] == 'X') & pd.notna(df['col2'])]`
- `df1[df1.column.isin(df2.column)]`

Selecting rows and columns

- `df.iloc[:, [0, 3, 6]]`
- `df.loc['index', 'column']`
- `df.loc[:, ('column_L0', 'column_L1')]`
- `df.drop('index')`
- `df.drop(columns = 'column')`
- `df['column'].drop_duplicates()`

Outline

Introduction
Data creation, input, and output
Data exploration
Summary statistics
Subset
Sorting
Data manipulations

Exercise

Sorting

- `df = df.sort_values('column') # sort rows by values of column`
- `df.sort_values('column', inplace = True)`

- `df = df.sort_index() # sort index labels`
- `df = df.sort_index(axis = 1) # sort column labels`

Outline

Introduction
Data creation, input, and output
Data exploration
Summary statistics
Subset
Sorting
Data manipulations

Exercise

Data manipulations

- `df['col'] = pd.to_numeric(df['col'])`
- `df['col'] = df['col'].astype(int)`
- `df['col'].to_list()`

- `df['column'].replace('a', 'b', inplace = True)`
- `df.column = df.column.fillna('notna')`

- `df = pd.concat([df1, df2], axis = 1)`
- `df = df1.merge(df2, how = 'inner', left_on = 'column1', right_on = 'column2')`

Outline

Introduction
Data creation, input, and output
Data exploration
Summary statistics
Subset
Sorting
Data manipulations

Exercise (live coding)

Exercise

- Import the data: shelter_count.csv
- Test some of the methods on slide 12 to explore the data
- Confirm that there are no missing values
- Delete the ‘total’ rows and columns, and then recalculate the totals
- Select the gross live outcomes from the grand total and sort the values in descending order
- Test the to_string and to_list methods
- Calculate net intakes and net live outcomes (gross values minus transfers between agencies), and then calculate the totals of both cats and dogs
- Merge the two series and export the data

Icon made by Freepik from Flaticon