Final Takeaway

Xavier Morera

Helping developers create amazing applications

@xmorera / www.xaviermorera.com / www.bigdatainc.org

Takeaway

Different types of files available

- Text, music, images, data...

Text files can hold different types of information

- Plain text, numeric data, or tabular data

Flat file

- Contains data in plain format
- Simple, two-dimensional structure

For example

- CSV, TSV, fixed-width

Takeaway

Parse flat files

- Manually
- CSV Reader module
- Pandas and Numpy

Read and process files manually

- Open, read, evaluate, print, and close
- Not ideal and error prone

Reader module automates the loading part

Takeaway

Numpy is a widely used library for scientific computing

- Used as a base for other libraries

ndarray is Numpy's base object

- Homogeneous multidimensional array

Contains multiple built-in functionalities

- Arithmetic, matrix, aggregation, broadcasting, reshaping, concatenation, splitting, transposing, and masking

Takeaway

Can be initialized in several ways

In memory

- List, tuple, nested list

From a file

- loadtxt and genfromtxt
- recfromcsv and recfromtxt

Takeaway

Pandas is an open-source library

- Data manipulation and analysis

Read and write in multiple formats

- Tabular data (CSV, TSV)
- Excel, JSON, XML, SQL databases...

Takeaway

First-class citizen is the dataframe

- Internally uses Numpy ndarray

Two-dimensional table-like data structure

Consists of rows and columns

- Row represents observation or data point
- Column represents a variable or feature

Like a spreadsheet or SQL table

Takeaway

Use read_csv to load tabular data

- CSV files

Many parameters to control how files are loaded and displayed

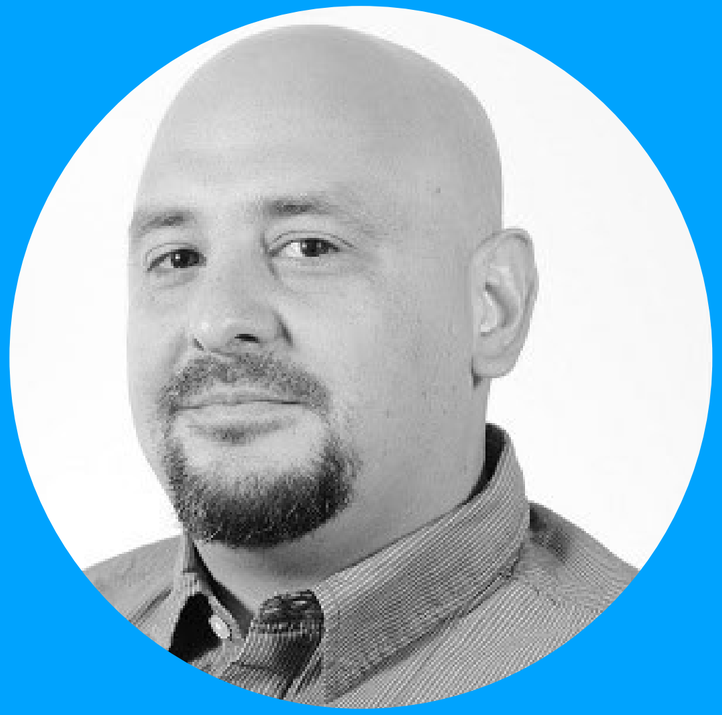
- head, nrows, skiprows, skipfooter, usecols, names, header, columns...

Use read_table to load TSV files

Can read and write to other formats

- Covered in separate trainings

Thanks for watching!



@xmorera

www.xaviermorera.com

“What you learn is yours for life.”