What is pandas?

INTRODUCTION TO DATA SCIENCE IN PYTHON

Fast, powerful, flexible and ***easy to use*** open source ***data analysis*** and ***manipulation*** tool, built on top of the Python programming language.

Pandas is well suited for many different kinds of data

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet.
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data with row and column labels.
- Any other form of observational/ statistical data sets.

Excel vs Pandas (Python)

Pandas

- Extremely fast and efficient.
- No real limit and handles millions of data points seamlessly
- Pandas can handle over 15 different formats and switch between them with ease. eg csv,SQL, json
- Advanced statistics and machine learning capabilities.
- Advanced data visualization capabilities.
- It’s easier for others to reproduce and audit your work.

Revenue

- In Excel, once you exceed 10,000 rows, it starts to slow down
- Excel caps a single spreadsheet at 1,048,576 rows exactly.
- You would have to spend time converting file formats before importing them,

Comparison with SQL

https://pandas.pydata.org/pandas-docs/stable/getting_started/comparison/comparison_with_sql.html

Installation

Working with conda?

- conda install pandas

Working with pip?

- pip install pandas

Additional Resources

Pandas Documentation

https://pandas.pydata.org/docs/getting_started/intro_tutorials/index.html

Modern Pandas

https://github.com/TomAugspurger/effective-pandas