What is Pandas?

pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive.

It aims to be the fundamental high-level building block for doing practical, **real world** data analysis in Python. Additionally, it has the broader goal of becoming the **most powerful and flexible open source data analysis / manipulation tool available in any language**.

It is already well on its way toward this goal.

Why Pandas?

Pandas surpasses excel in a few important ways:

- Integration in the python ecosystem. A single (reproducible!) script for your data retrieval, cleaning, plotting, and (advanced) analysis.
- Natively understands familiar formats (xls and csv) as well as "heftier" ones like json and sql (next week).
- More powerful for cleaning, slicing, and merging multiple datasets.
- Flexible and scriptable plotting (matplotlib).
- Far more powerful statistical modeling (statsmodels).
- Beyond this course – pandas is built on numpy, and is computationally very efficient.

Core Concepts

Two datatypes: pandas.Series (columns) and pandas.DataFrame (tables).

Three big ideas:

1. **masking** rows and selecting columns: selecting sets of rows or columns from a dataset.
   - Select columns: `df[column_name]` or `df[column_list]`
   - Masking rows: `df[mask]`
   - Both simultaneously: `df.loc[mask, column_list]`.
2. **merging** or **joining** data: assembling one dataset from multiple files.
   - Join by index: `df1.join(df2)`
   - Merge on columns: `pandas.merge(df1, df2)`
3. **aggregating**: perform operations on subsets of the data in one pass.
   - Group by: `df.groupby(value).mean()`, `.sum()`, `.median`, etc.