Pandas Theory Document

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 easily and intuitively. 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.

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 (homogeneously typed or heterogeneous) with row and column labels
- Any other form of observational / statistical data sets. The data actually need not be labeled at all to be placed into a pandas data structure

The two primary data structures of pandas, `Series (1-dimensional)` and `DataFrame (2-dimensional)`, handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering. For R users, DataFrame provides everything that R's `data.frame` provides and much more. pandas is built on top of `NumPy` and is intended to integrate well within a scientific computing environment with many other 3rd party libraries.

Here are just a few of the things that pandas does well:

- Easy handling of missing data (represented as NaN) in floating point as well as non-floating point data
- Size `mutability`: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, DataFrame, etc. automatically align the data for you in computations
- Powerful, flexible group by functionality to perform `split-apply-combine` operations on data sets, for both aggregating and transforming data
- Make it easy to `convert ragged, differently-indexed data` in other Python and NumPy data structures into DataFrame objects
- Intelligent `label-based slicing`, `fancy indexing`, and `subsetting of large data sets`

- Intuitive `merging` and `joining` data sets
- Flexible `reshaping` and pivoting of data sets
- `Hierarchical labelling` of axes (possible to have multiple labels per tick)
- Robust IO tools for loading data from `flat files (CSV and delimited)`, Excel files, databases, and saving / loading data from the ultrafast HDF5 format
- `Time series-specific functionality`: date range generation and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging, etc.

Many of these principles are here to address the shortcomings frequently experienced using other languages / scientific research environments. For data scientists, working with data is typically divided into multiple stages: munging and cleaning data, analysing / modelling it, then organising the results of the analysis into a form suitable for plotting or tabular display. pandas is the ideal tool for all of these tasks.