

pandas is a high-performance module that provides a comprehensive set of structures for working with data. pandas excels at handling structured data, such as data sets containing many variables, working with missing values and merging across multiple data sets.

While extremely useful, pandas is not an essential component of the Python scientific stack unlike NumPy, SciPy or matplotlib, and so while pandas doesn't make data analysis possible in Python, it makes it much easier. pandas also provides high-performance, robust methods for importing from and exporting to a wide range of formats.

pandas provides a set of data structures which include Series, DataFrames and Panels. Series are 1-dimensional arrays. DataFrames are collections of Series and so are 2-dimensional, and Panels are collections of DataFrames, and so are 3-dimensional. Note that the Panel type is not covered in this chapter.

head() shows the first 5 rows of a series, and tail() shows the last 5 rows. An optional argument can be used to return a different number of entries, as in head(10).

0.0.1 isnull and notnull

isnull() returns a Series with the same indices containing Boolean values indicating True for null values which include NaN and None, among others.

notnull() returns the negation of isnull() – that is, True for non-null values, and False otherwise.

0.0.2 ix

ix is the indexing function and s.ix[0:2] is the same as s[0:2]. ix is more useful for DataFrames.

0.0.3 plot

plot is the main plotting method, and by default will produce a line graph of the data in a DataFrame. Calling plot on a DataFrame will plot all series using different colors and generate a legend. A number of keyword arguments are available to affect the contents and appearance of the plot.

- style, a list of matplotlib styles, one for each series plotted. A dictionary using column names as keys and the line styles as values allows for further customization.
- title, a string containing the figure title.
- subplots, a Boolean indicating whether to plot using one subplot per series (True). The default is False.
- legend, a Boolean indicating whether to show a legend

Data Analysis with Python

- `secondary_y`, a Boolean indicating whether to plot a series on a secondary set of axis values. See the example below.
- `ax`, a matplotlib axis object to use for the plot. If no axis is provided, then a new axis is created.
- `kind`, a string, one of: