pandas Dataframes:
A Myriad of Accessors

DA 02505 Mini-Talk

Graham Umlauf August 8 2018

Introduction to Dataframes

- In pandas, a dataframe is a set of heterogeneous data organized by columns and rows.
- Each column has a label, which can be of most variable types.
- Rows are defined by an index, usually an integer.
- Equivalent to an R dataset.

Dataframe Info at a Glance

- `df.head()`
  - Returns the data in the top few rows, as well as the column labels. Often truncated by width.
  - df.tail() returns the bottom few rows.
- `df.columns`
  - Returns a list of the column names.
- `df.index`
  - Returns a list of row indexes.

Dataframe Accessors

- A dataframe is useless unless the data can be accessed.
- While single-purpose programs mostly allow for very easy and straightforward access, pandas also provides many accessor methods that enable more modular access, even for dataframes for which some information may not be known about ahead of time.

Accessing Data, the Intuitive Way

- `df['A']`
  - Returns the data in column A as a list.
  - `df['A'][3]` returns the fourth data point in column A.
  - WARNING: This method returns a copy of the data. It cannot be used to assign values to the dataframe.

Accessing by Location

- `df.loc[row, column]`
  - Returns the data at the specified row and column. NOTE: row comes first, whereas in direct access, column comes first.
  - `x:y` can be used to define a range.
  - `:` alone returns the entire range of that row or column.
  - Returns data by reference, so the dataframe can be modified via `df.loc[x, 'A'] = 5.0`

Accessing By Position

- `df.iloc[ii, jj]`
  - Accesses data by integer index regardless of the datatype or labels of the rows and columns.
  - Row comes first, as with `loc`.
  - Returns a reference, allowing assignment as with `loc`.

And More:

- Optimized Scalar Access: `df.at[x, 'A']` and `df.iat[ii,jj]`
  - Similar to `loc` and `iloc`, but optimized for rapid access of individual data points.
- Boolean indexing: `df[df.A > 1.0]`
  - Uses a boolean list to return only the rows corresponding to `True`
  - Similar in function to Matlab logical indexing