



Introduction to Data Science

GIRI NARASIMHAN, SCIS, FIU

Momentos Survey

- ▶ Survey Consent

<https://users.cs.fiu.edu/~giri/Momentos/MomentosConsentForm.pdf>

- ▶ Register

Course Code: 295MFN

- ▶ Survey link

tinyurl.com/premomentospre

Personal Code: XXXX

Case History

- ▶ MovieLens1M.ipynb

NumPy: numerical computing packages

- ▶ Fast and efficient multidimensional array object ***ndarray***
- ▶ Functions for **element-wise array computations** and array operations
- ▶ Tools for **reading and writing** array-based data sets to disk
- ▶ **Linear algebra operations, Fourier transform, and random number generation**
- ▶ Tools for integrating connecting C, C++, and Fortran code to Python
- ▶ NumPy arrays are **more efficient** way of storing and manipulating data and better for passing between algorithms. Libraries in C or Fortran can operate on NumPy arrays without copying any data.

Pandas: package for structured data

- ▶ **DataFrame**: more general than R's data.frame
- ▶ Combines NumPy arrays with manipulations similar to spreadsheets and relational databases
- ▶ Sophisticated **indexing** facilities
- ▶ **Reshape, slice** and **dice, aggregations, subselections**, etc.
- ▶ **Time series** processing functionality

pandas DataFrames

Table 5-1. Possible data inputs to DataFrame constructor

Type	Notes
2D ndarray	A matrix of data, passing optional row and column labels
dict of arrays, lists, or tuples	Each sequence becomes a column in the DataFrame. All sequences must be the same length.
NumPy structured/record array	Treated as the “dict of arrays” case
dict of Series	Each value becomes a column. Indexes from each Series are unioned together to form the result’s row index if no explicit index is passed.
dict of dicts	Each inner dict becomes a column. Keys are unioned to form the row index as in the “dict of Series” case.
list of dicts or Series	Each item becomes a row in the DataFrame. Union of dict keys or Series indexes become the DataFrame’s column labels
List of lists or tuples	Treated as the “2D ndarray” case
Another DataFrame	The DataFrame’s indexes are used unless different ones are passed
NumPy MaskedArray	Like the “2D ndarray” case except masked values become NA/missing in the DataFrame result

Index objects

Table 5-2. Main Index objects in pandas

Class	Description
Index	The most general Index object, representing axis labels in a NumPy array of Python objects.
Int64Index	Specialized Index for integer values.
MultiIndex	"Hierarchical" index object representing multiple levels of indexing on a single axis. Can be thought of as similar to an array of tuples.
DatetimeIndex	Stores nanosecond timestamps (represented using NumPy's <code>datetime64</code> dtype).
PeriodIndex	Specialized Index for Period data (timespans).

More on Index

Table 5-3. Index methods and properties

Method	Description
<code>append</code>	Concatenate with additional Index objects, producing a new Index
<code>diff</code>	Compute set difference as an Index
<code>intersection</code>	Compute set intersection
<code>union</code>	Compute set union
<code>isin</code>	Compute boolean array indicating whether each value is contained in the passed collection
<code>delete</code>	Compute new Index with element at index <code>i</code> deleted
<code>drop</code>	Compute new index by deleting passed values
<code>insert</code>	Compute new Index by inserting element at index <code>i</code>
<code>is_monotonic</code>	Returns True if each element is greater than or equal to the previous element
<code>is_unique</code>	Returns True if the Index has no duplicate values
<code>unique</code>	Compute the array of unique values in the Index

SciPy: scientific computing packages

- ▶ **scipy.integrate**: numerical integration routines and differential equation solvers
- ▶ **scipy.linalg**: linear algebra, matrix decompositions extending beyond numpy.linalg.
- ▶ **scipy.optimize**: function optimizers (minimizers) and root finding algorithms
- ▶ **scipy.signal**: signal processing tools
- ▶ **scipy.sparse**: sparse matrices and sparse linear system solvers
- ▶ **scipy.special**: wrapper around SPECFUN, a Fortran library implementing many common mathematical functions, such as the gamma function
- ▶ **scipy.stats**: standard continuous and discrete probability distributions (density functions, samplers, continuous distribution functions), various statistical tests, and more descriptive statistics
- ▶ **scipy.weave**: tool for using inline C++ code to accelerate array computations

matplotlib: for visualization

- ▶ **Matplotlib:** Python library for publication-quality visualizations
- ▶ Creator: John D. Hunter, but maintained by team of developers
- ▶ Can be used in **notebooks** with **interactive** features; zoom in on section of plot and pan around using the toolbar in plot window.

Two kinds of data structures

- ▶ Structured
 - ❑ Lists: Arrays, Tables and Spreadsheets
 - ❑ Strings
 - ❑ Matrices: Images
 - ❑ Dictionaries: for Associations
 - (Key, Value) Pairs
 - ❑ Time Series & Trajectories
 - Audio, Video
- ▶ Unstructured e.g., text
- ▶ Maps: (functions, data) pair