UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

3250 Foundations of Data Science

Module 4: Pandas

Course Plan

Module Titles
Module 1 – Introduction to Data Science
Module 2 – Introduction to Python
Module 3 – NumPy
Current Focus: Module 4 – Pandas
Module 5 – Data Collection and Cleaning
Module 6 – Descriptive Statistics and Visualization
Module 7 – Workshop (No Content)
Module 8 – Time Series
Module 9 – Introduction to Regression and Classification
Module 10 – Databases and SQL
Module 11 – Data Privacy and Security
Module 12 – Term Project Presentations (no content)

Learning Outcomes for this Module

- Further build your Python skills
- Use Pandas data analysis libraries to organize and summarize data

Topics for this Module

- **4.1** Pandas: the Python data analysis package
- **4.2** Class Exercises
- **4.3** Resources and Homework

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 4 – Section 1

Pandas

Pandas

- Data analysis package created by Wes McKinney
- Brings the equivalent of the R Data Frame to Python
- Powerful capabilities for working with time series data
- Automatic data alignment
- Flexible handling of missing data
- Relational operations
- Support for categorical variables

Series

- One-dimensional array of data with a one-dimensional array of labels called the *index*
- Usually of a single type but can be heterogeneous
- We'll come back to it in the module on Time Series

DataFrame

- A tabular data structure with labelled columns and rows
- Used for manipulating and analyzing data
- Exhibits size mutability allowing rows and columns to be added and deleted
- Similar to a relational table but heterogeneously-typed

DataFrame

	Age	Height	Weight
0	8	128	27.5
1	10	138.9	34.5
2	16	157.3	91.1
3	6	116.6	21.4
4	14	159.2	54.4

Creating a DataFrame

- Can be created from:
  - Dict of 1-D structures (ndarrays, lists, dicts, tuples or Series)
  - List of 1-D structures
  - 2-D numpy ndarray
  - Structured or record ndarray
  - A Series
  - Another DataFrame

Creating a DataFrame from a 2-D List

```
import pandas as pd
df = DataFrame(
            data=[
                  [8, 128, 27.5],
                  [10, 138.9, 34.5],
                  [16, 157.3, 91.1],
                  [6, 116.6, 21.4],
                  [14, 159.2, 54.4]
            ],
            columns=[“Age”, “Height”, “Weight”]
      )
```

Indexing for DataFrames

- Use the method `.ix` to select rows
- Example:
  `df.ix[0]`
  `df.ix['Toronto']`
- Use either of these forms for columns:
  `df['Age']`
  `df.Age`

Loading/Saving DataFrames

- `import pandas as pd`
- CSV: `pd.read_csv(), pd.to_csv()`
- Excel: `pd.read_excel(), pd.to_excel()`
- Relational tables: `pd.read_sql(), pd.to_sql()`
- Pandas SQL queries

Basic Statistical Functions

- Mean
  - pandas.DataFrame.mean
- Median
  - pandas.DataFrame.median
- Standard Deviation
  - pandas.DataFrame.std
- Sum
  - pandas.DataFrame.sum

Hierarchical Indexes in DataFrames

- DataFrames can have a hierarchy of indexes, e.g.

```
df = DataFrame(
          data=[4, 7, 2, 5, 6],
          columns=[“Data”],
          index=
                [“a”, “a”, “b”, “b”, “a”],
                [“x”, “y”, “x”, “y”, “x”]
          ])
```

		Data
a	x	4
a	y	7
b	x	2
b	y	5
a	x	6

Aggregation and Grouping

- Pandas has “slice and dice” operations for DataFrames
  - Split into pieces by key
  - Apply functions to each column
  - Apply functions to groups
  - Compute pivot tables
  - Calculate common statistics by group

GroupBy

- Works by split-apply-combine
- Identify a grouping with `.groupby()` e.g. `df.groupby('key1')`
- This creates a `GroupBy` object but doesn't actually do the split-apply-combine yet
- Iterate over the groups using `group in df.groupby('key1')`

Pivot Tables

- DataFrame has a `.pivot_table()` method that makes it easy to select rows and columns of a hierarchically indexed DataFrame and get automatic subtotals by group

Module 4 – Section 2

Class Exercises

Getting Started with Pandas

- Chapter 5 of "Python for Data Analysis" book, Getting Started with Pandas

- Download the code (or clone the repository) and work through Chapter 5

Analyzing 311 Calls in New York

- Pandas Cookbook, Julia Evans: Chapters 2 and 3

Linear Regression with Python

- Linear Regression with Python blog article by Connor Johnson

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 4 – Section 3

Resources and Homework

Resources

- Intro to Data Structures:
- Open Data with IPython and Pandas:
- Some tricks for better looking Pandas tables:

Next Class

- Gathering and Preparing Data
  - Types of data
  - Reading and saving data
  - Cleaning up problem data using Pandas
  - Handling missing data
  - Getting data from the web

Follow us on social

Join the conversation with us online:

facebook.com/uoftscs

@uoftscs

linkedin.com/company/university-of-toronto-school-of-continuing-studies

@uoftscs

Any questions?

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Thank You

Thank you for choosing the University of Toronto School of Continuing Studies