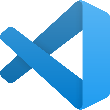
**CS506 Programming for Computing**

**HOP06C – Getting High-Performance with pandas**

06/09/2020 Created by Apiwat Chuaphan

11/08/2020 Revised by Kim Nguyen

Center for Information Assurance (CIAE) @City University of Seattle (CityU)

 A close up of a screen

Description automatically generated A close up of a sign

Description automatically generated

**Before You Start**

* The directory path shown in screenshots may be different from yours.
* Some steps might not be explained in the tutorial.  If you are not sure what to do:
  + Consult the resources listed below.
  + If you cannot solve the problem after a few tries, ask a TA for help.

**Learning Outcomes**

* Understand pandas object
* Learn the basic functionality

**Resources**

* Tutorialpoint
* Pandas User Guide: <https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html>
* McIntire, G., Martin, B., & Washington, L. Pandas A Complete Introduction. Retrieved from <https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/>

**What is** **pandas?**

pandas (derived from the word Panel Data – an Econometrics from Multidimensional data – Tutorialspoint) is a powerful, open source Python library providing high-performance, easy-to-use data structures for data analysis, manipulation, and visualization.

Features of Pandas

* Fast and efficient DataFrame object
* Tools for loading data into in-memory data objects from different file formats.
* Label-based slicing, indexing and subsetting of large data sets.
* Columns from a data structure can be deleted or inserted.
* Group by data for aggregation and transformations.
* High performance merging and joining of data.

**The First Step:**

We need to install pandas.

1. In Visual Studio Code, open the private repository generated when you accepted the HOP06 assignment (If you cannot find that repository in your machine, you might have not cloned the repo, if so, please do before proceeding).
2. Open terminal (Control + `) in VS Code, then execute the command:

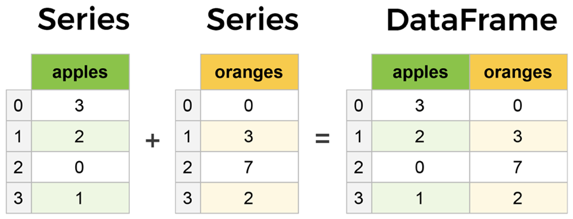
>>> pip install pandas

**Core Components**

The two primary data structures of pandas, **Series** and **DataFrame**.

* **Series** – 1D labeled homogeneously-typed array
* **DataFrame** – General 2D labeled, size-mutable tabular structure with potentially heterogeneously-typed column

A **Series** is essentially a column, and a **DataFrame** is a multi-dimensional table made up of a collection of Series.



Open Jupyter Notebook:

1. Under module folder, create a new file called **pandas\_object.ipynb** and simply click on the file to open notebook.
2. Type the following into the file just created. Run selected cell to see each result.

A screenshot of a cell phone

Description automatically generated

1. We just created Series object with pandas, next we will create DataFrame, which is a collection of Series.

A screenshot of a cell phone

Description automatically generated

Each (key, value) item in data corresponds to a column in the resulting DataFrame.

1. The previous DataFrame, index was given at the creation by default, but we can create our own labels as the following example.

A screenshot of a cell phone screen with text

Description automatically generated

We gave string names as index replacement, so it is more convenient to l**oc**ate the data by name. For example:

buyer.loc['Tom']

Output:

A screenshot of a cell phone

Description automatically generated

That’s the basic pandas, there is a lot more from pandas that you can do. Learn more here: <https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html>