Advanced Data Analysis - Python

M. Fuat Kına

fuatkina@gmail.com

fuat.kina@marmara.edu.tr

mkina18@ku.edu.tr

mkina@ku.edu.tr

Course objectives

- Learn Python 3 syntax
- Understand basic programming concepts
- Understand advanced data analysis problems and the needed tools to solve them
- Establish a basic understanding of machine learning concepts and algorithms

Why Python?

- Great for beginners and good for advanced use
  - Easily readable code
  - Online resources
- Widely used, especially in scientific computing
- Powerful
  - Advanced data analysis techniques
  - Machine learning modules
- Open-source
- Alternatives for data analysis: R, STATA, SPSS, GIS programs (ArcGIS, QGIS, Geoda) etc.

Course content

- Python basics (data types, lists, sets, dictionaries, basic operations, if statements, functions, and loops)
- Data collection (web scrapping)
- Working with data, creation and manipulation (numpy, matplotlib, pandas)
- Advanced data analysis techniques
  - OLS
  - Spatial statistics
  - Bayesian statistics
  - Machine learning

Course materials

- VanderPlas, Jake. 2016. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/
- Shaw, Zed A. 2017. Learn Python 3 the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code (Zed Shaw's Hard Way Series). 1st Edition. Addison-Wesley. Available at: https://learnpythonthehardway.org/python3/
- The Official Documentation for Python. Available at: https://docs.python.org/3/
- Weekly readings and links for some online sources and Python scripts
- Two more books:
  - Angrist, J.D., & Pischke, J.S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion, Princeton: Princeton University Press.
  - LeSage, J., & Pace, R. K. (2009). Introduction to Spatial Econometrics. Chapman and Hall/CRC.

Grading

- Assignments: %45
  - Assignment 6 will be graded out of 10 points, while all other assignments will be out of 7 points.
- Midterm report: %5
- Final project: %30
  - Class presentation, written report (approximately 2000 words), a python script that you used, and a Github repo which includes all. You are expected to submit only the Github link.
- Attendance: %20
- **Prepare yourself before class (must)**

Assignmets

- *Assignment 1*: Programming.
- *Assignment 2*: Pandas.
- *Assignment 3*: Scraping.
- *Assignment 4*: Statistics.
- *Assignment 5*: Spatial statistics.
- *Assignment 6:* Machine Learning. For this last assignment, you are expected to develop an analytical framework through machine learning, including preprocessing, model selection, training and testing, hyperparameter tuning, and presenting a group of outputs. This assignment is designed like a shared task. The higher scores will be awarded. Details for the task will be announced.

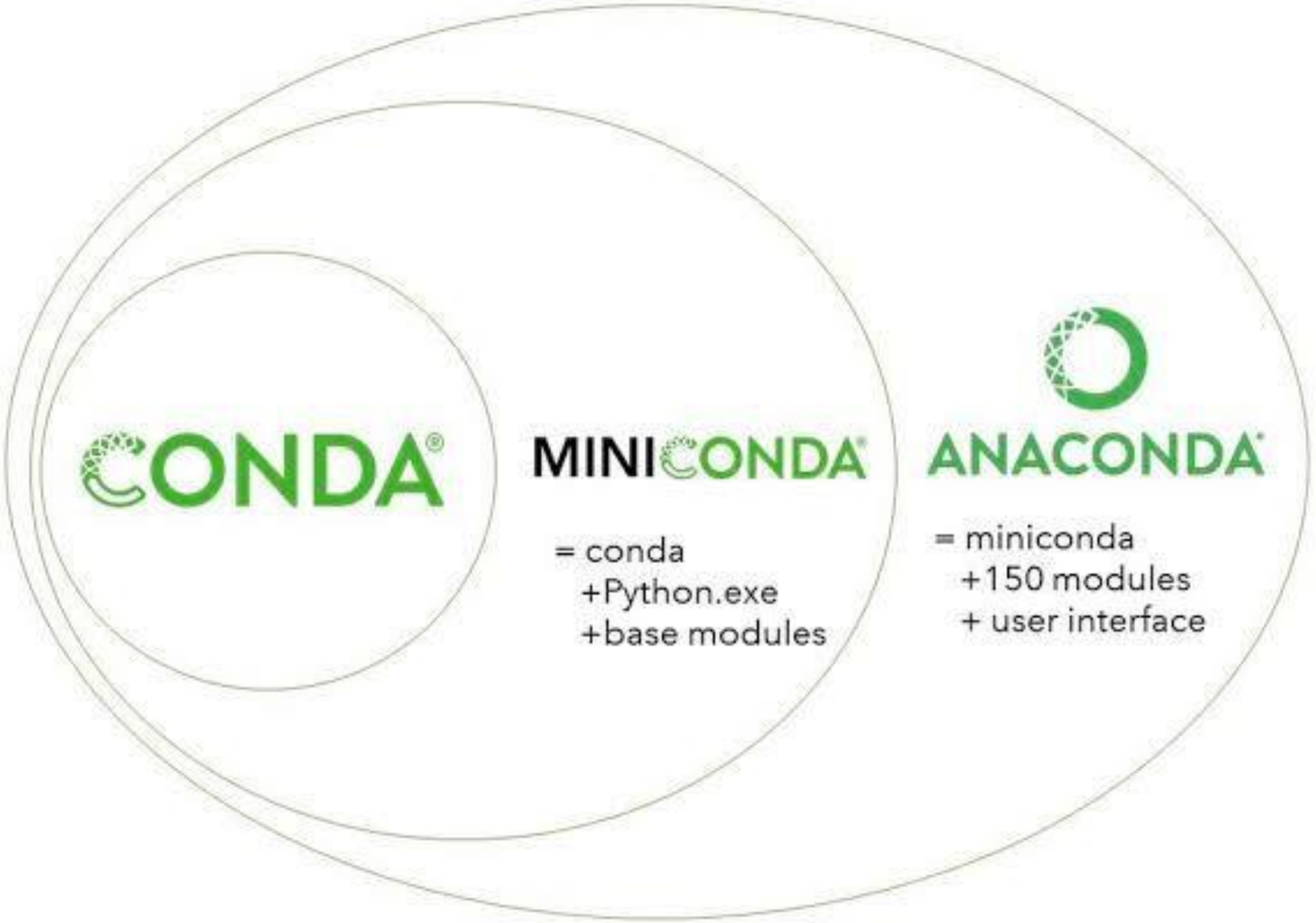
Let’s look at the syllabus

Let’s look at installation guideline!

You can run python without downloading anything

https://colab.research.google.com/

Conda, miniconda, anaconda



Conda environments

- What is an environment?
  - A conda environment is a directory that contains a specific collection of conda packages that you have installed. For example, you may have one environment with NumPy 1.7 and its dependencies, and another environment with NumPy 1.6 for legacy testing.
  - https://docs.conda.io/projects/conda/en/latest/user-guide/concepts/environments.html#:~:text=A%20conda%20environment%20is%20a,NumPy%201.6%20for%20legacy%20testing.
- How to manage an environment?
  - https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html#id2

How to use Python?

- Go for the installation guidelines
- Download the Python 3
- Install Anaconda
- Inspect Google Colab
- Play with conda
- Work on Shaw, Exercises 0-40