# Faculty notes - Working with data in Python

#### Overview

This 4-hour workshop takes learners through the basics of programming in Python via the Jupyter Lab interface and culminates with exploration and visualization of real-world bicycle count data from the City of Toronto. This material focuses on using the package pandas for working with spreadsheet-type data and the packages matplotlib and seaborn for data visualization. The material is designed to be delivered as a participatory live-coding workshop where the instructor projects their computer screen while coding and learners follow along on their own computers. This material is based on workshops hosted by UofT Coders, inspired by the Data Carpentry Ecology Python lesson.

- Prerequisites: This material assumes no background knowledge of programming.
- Target audience: Undergraduates, graduate students, faculty, or staff in any discipline.

# Learning objectives

## Part 1: introduction to programming in Python

- Overview of the capabilities of Python and how to use JupyterLab for exploratory data analyses.
- Learn about some differences between Python and Excel.
- Learn basic Python commands.
- Learn about the Markdown syntax and how to use it within the Jupyter Notebook.

### Part 2: working with data in Python

- Describe what a data frame is
- Load external data from a .csv file into a data frame with pandas
- Summarize the contents of a data frame with pandas.
- Learn to use data frame attributes loc[], head(), info(), describe(), shape, columns, index.
- Understand the split-apply-combine concept for data analysis.
- Use groupby(), sum(), agg() and size() to apply this technique.

#### Part 3: visualizing data

- Produce scatter plots, line plots, and histograms using seaborn and matplotlib.
- Understand how to graphically explore relationships between variables.
- Apply grids for faceting in seaborn.
- Set universal plot settings.
- Use seaborn grids with matplotlib functions

### Lesson outline

- Communicating with computers (5 min)
  - Advantages of text-based communication (5 min)
  - Speaking Python (5 min)
  - Natural and formal languages (5 min)
- The Jupyter Notebook (10 min)
- Data analysis in Python (5 min)
  - Packages (5 min)
  - How to get help (5 min)

- Manipulating and analyzing data with pandas
  - Data set background (10 min)
  - What are data frames (15 min)
  - Data wrangling with pandas (40 min)
- Split-apply-combine techniques in pandas
  - Using sum() and mean() to summarize categorical data (20 min)
  - Using size() to summarize categorical data (10 min)
- Data visualization with matplotlib and seaborn (10 min)
  - Visualizing one quantitative variable with multiple categorical variables (40 min)
  - Visualizing the relationship of two quantitative variable with multiple categorical variables (40min)
  - Using any plotting function with seaborn grids (10 min)

## Data description

Parts 2 and 3 of this material use data from the City of Toronto Open Data Catalogue, a great resource with lots of publicly available data. The dataset used is counts of bicycles from the College St. bikelanes in September 2010 and September 2017. I cleaned the data and processed it into a long spreadsheet format that is used in this lesson. The cleaned data can be downloaded at this link: https://bit.ly/2Cs1Mq1 or https://gist.githubusercontent.com/mbonsma/be7482639d7a2d5cfc52505aadb9b53e/raw/1f68fce4a127fdd3b2313728dd84cf21e86e7df3/college\_spadina\_2010\_2017.csv

## Challenges to address

Participatory live-coding is suitable for small to medium-sized groups of learners. This workshop was presented to a group of about 35 people, and I would recommend a group no larger than 40. It is very helpful to have several helpers present that can move around the room and help learners debug while the workshop progresses. I recommend using a sticky note system for monitoring the pace and for letting learners flag when they need help.