A Practical Introduction to Data Science with Python

AccessPay/MMU

Luciano Gerber

Senior Lecturer
Department of Computing and Mathematics
Faculty of Science and Engineering
Manchester Metropolitan University

June 2019

Preamble

Outline

- a hands-on introduction to Data Science with Python, including data manipulation, visualisation, and exploration.
- live, interactive demo
- topics include. . .
  - a feel for the Data Science process
  - Data Manipulation (e.g., loading, inspecting, subsetting) (`pandas`)
  - Data Exploration, Transformation, Cleaning (`pandas`, `seaborn`)
  - Data Merging, Grouping and Aggregation (`pandas`)
  - Data Visualisation (`matplotlib`, `seaborn`)
- taught to MSc students over 6 weeks of 5-hour sessions!

Setup

- if people want to follow along, try the code and attempt some exercises. . .
  - https://github.com/gerberl/accesspay-mmu-intro-ds
  - regardless of platform (e.g., Windows, Linux, Mac), the Anaconda distribution is probably the best way forward for installing and managing your Python installation with required modules for the Data Science ecosystem.
  - in Windows, it is probably best to invoke `ipython` via the Anaconda shell. In Linux and Mac, use the terminal for both `ipython` and `jupyter`.
  - you will need a programming text editor, such as `Sublime Text`, `Atom`, `emacs`, among others. IDEs (e.g., `Spyder`, `PyCharm`, `Visual Studio Code`) will do too.

Mindset

- **exploratory** and **interactive** first.
- `ipython` REPL and `jupyter` notebooks as the environments.
- programming text editor (e.g., `sublime`); copy-and-paste.
- documentation: integral to the process (**literate programming**).
- automation? production? Make scripts from the output of the exploratory stages.

Resources

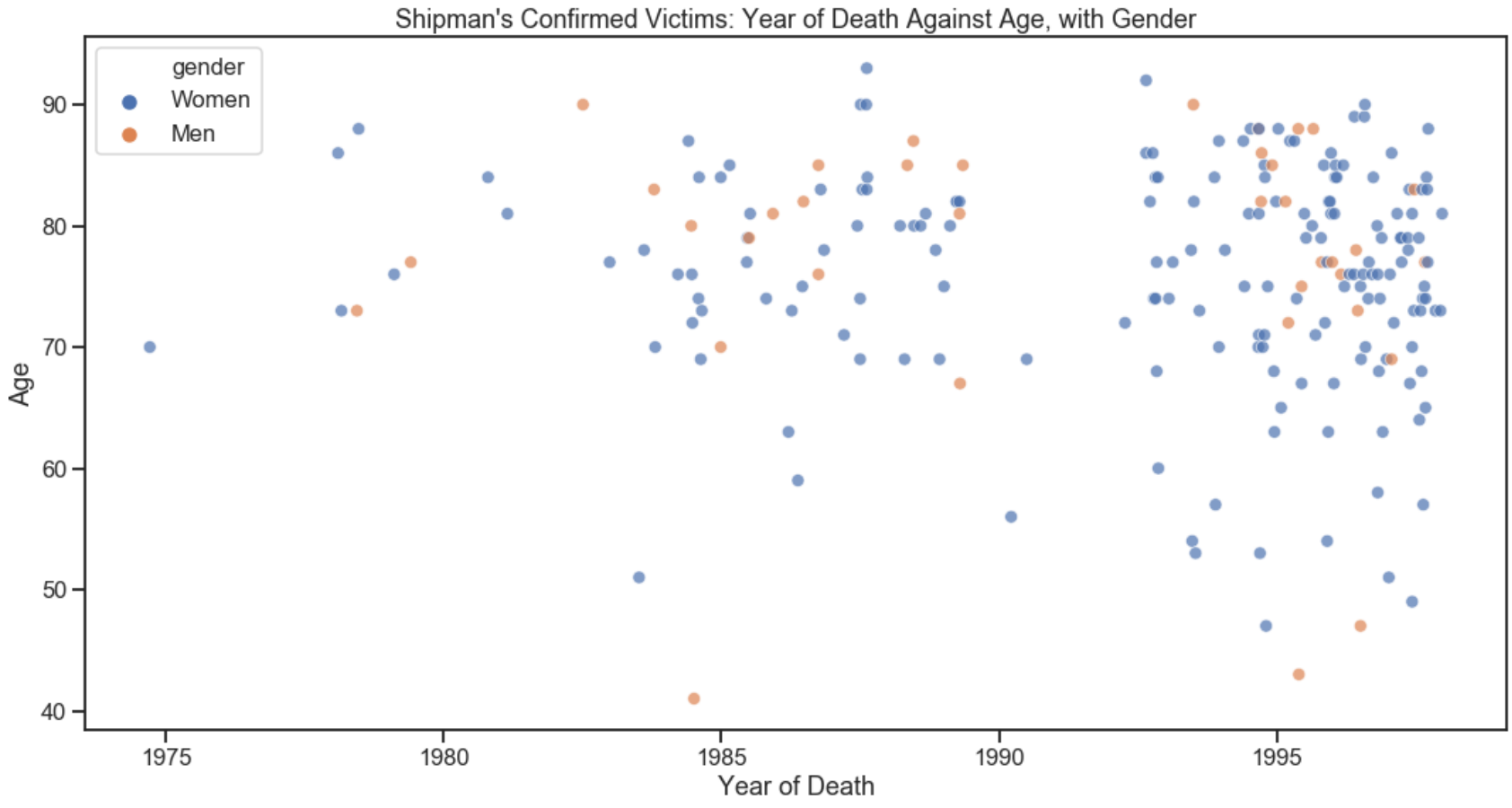
- some familiarity with python essentials is helpful. As a pre-requisite, I suggested that people could try the following before the workshop:
  - DataCamp's interactive tutorial on Python for Data Science: https://www.datacamp.com/courses/intro-to-python-for-data-science
- some other useful, companion resources are:
  - A whirlwind tour of python: https://jakevdp.github.io/WhirlwindTourOfPython/
  - Python Data Science Handbook: https://jakevdp.github.io/PythonDataScienceHandbook/.

Diving In: A Quick Example

Background

- a little macabre, but an interesting example from The Art of Statistics:
  - what could one learn from Harold Shipman's serial killings?
  - could it have been detected/prevented (with Data Science)?

Victims by Year of Death, Age, and Gender



Victims by Time of Day, Compared to a Typical Physician