Data Science 1

Tobin Driscoll

1/1/23

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

Resources

Useful guides

- pandas user guide, pandas cheat sheet, pandas long summary
- seaborn tutorial, seaborn cheat sheet
- scikit-learn user guide, sklearn cheat sheet
- numpy cheat sheet

Data sources

Here are places around the web with data available for download.

Packaged

These feature datasets that are essentially already packaged as CSV or Excel files, plus descriptions.

- Five Thirty-Eight Data used to support the site's journalism, mainly in politics and sports.
- Delaware Open Data Publicly available data from the state government.
- Kaggle Long-time host of data science competitions. The formal competitions are well-curated, but user contributions vary widely.
- UCI Machine Learning Repository Long-standing site for datasets that have been used extensively in machine learning research, but also recent contributions.
- Open ML Sort of abandoned years ago, but lots of eclectic datasets remain.

- IMDB Datasets Information about movies and TVs. (Big files!)
- Stanford Network Analysis Project Datasets presented as networks.

Open-ended

These require you to navigate an interface to select data from a large pool. Typically, you can make selections, preview the dataset, and then download in CSV or Excel format.

- U.S. Census Bureau Tons of demographic data about the U.S.
- Data.gov Home for all open U.S. government data.
- UNICEF Portal Worldwide data about child welfare.
- World Bank Focuses on economic and development data.
- World Health Organization Information on health and disease.

Search engines

These point to a lot of other resources.

- Google Dataset Search
- Registry of Open Data on AWS Access to datasets used by governments and researchers that happen to be stored on Amazon's servers. Skewed toward large sets, and a bit of a grab-bag.

Glossary

A much more exhaustive glossary can be found here.

Git

- **Git** Protocol for maintaining the entire file history of a project, including all versions and author attributions.
- **repository** Collection of files needed to record the history of a git project.
- **GitHub** Website that hosts git repositories created by private users, along with tools to help inspect and manage them.
- **commit** Collection of particular changes to the repository made by an individual and given a message.
- **stage** Temporary designation of locally modified files to be added to the next commit.
- **merge** Automatic union of non-conflicting commits from different sources.
- **conflict** Disagreement between repository versions that requires human intervention to resolve.
- **push** Sending one or more commits from a local repository to a remote repository.
- **pull** Receiving and merging all commits from a remote repository that are unknown to the local repository.

Notebooks

- **notebook** Self-contained collection of text, math, code, output, and graphics.
- **kernel** Back-end that executes code from and returns output to the notebook.
- **cell** Atomic unit of a notebook that contains one or more lines of text or code.
- **Markdown** Simplified syntax to put boldface, italics, and other formatting within text.
- **TeX/LaTeX** Language used to express mathematical notation within a notebook.
- **Jupyter** Popular format and system for interactive editing, execution, and export of notebooks.
- **Jupyter Lab** Layer over Jupyter notebook functionality to help manage notebooks and extensions.

Python

- **package** (or wheel) Collection of Python files distributed in a portable way to provide extra functionality.
- **numpy** Package of essential tools for numerical computation.
- **scipy** Package of tools useful in scientific and engineering computation.
- **database** Structured collection of data, usually with a formal interface for interaction with the data.
- **data frame** Tabular representation of a data set analogous to a spreadsheet, in which columns are observable quantities and rows are different observations of the quantities.
- **pandas** Package for working with data frames.
- **matplotlib** Package providing plot capabilities, modeled on MATLAB.
- **seaborn** Package built on matplotlib and providing commands to simplify creating plots from data frames.
- **scikit-learn** Package that provides capabilities for machine learning using a variety of methods.
- **tensorflow**, **keras**, **pytorch** Best-known packages for machine learning using neural networks.
- **Anaconda** Bundle of Python, most of the popular Python packages, Jupyter, and other useful tools, all within an optional point-and-click interface.

Editors/IDEs

- **Jupyter** Popular format and system for interactive editing, execution, and export of notebooks.
- **Jupyter Lab** Layer over Jupyter notebook functionality to help manage notebooks and extensions.
- **PyCharm** Feature-rich freemium development environment for Python, geared toward large, complex projects.
- **VS Code** Free all-purpose editor with many extensions for working closely with git, Github, Jupyter, and Python.
- **Spyder** Free development environment that somewhat resembles MATLAB.

- **Thonny** Bare-bones development environment intended to prioritize beginners.
- **Google Colab** Free cloud-based service for jumping into Jupyter notebooks without installing any software locally.