Introduction to Data Science

STAT 3255/5255 @ UConn

Jun Yan

1/17/23

Table of contents

Preface

The notets are a Quarto book; for details visit https://quarto.org/docs/books.

The notes are a joint effort of the instructor and the students in STAT 3255/5255, Spring 2023.

1 Introduction

1.1 What Is Data Science?

One widely accepted concept is the three pillars of data science: mathematics/statistics, computer science, and domain knowledge.

In her 2014 Presidential Address, Prof. Bin Yu, then President of the Institute of Mathematical Statistics, gave an interesting definition:

$$\text{Data Science} = \text{SDC}^3,$$

where S is Statistics, D is domain/science knowledge, and the three C's are computing, collaboration/teamwork, and communication to outsiders.

1.2 Expectations from This Course

- Proficiency in project management with Git.
- Proficiency in project report with Quarto.
- Hands-on experience with real-world data science project.
- Competency in using Python and its extensions for data science.
- Full grasp of the meaning of the results from data science algorithms.
- Basic understanding the principles of the data science methods.

1.3 Computing Environment

All setups are operating system dependent. As soon as possible, stay away from Windows. Otherwise, good luck (you will need it).

1.3.1 Command Line Interface

On Linux or MacOS, simply open a terminal.

On Windows, several options can be considered.

- Cygwin (with X): https://x.cygwin.com
- Git Bash: https://www.gitkraken.com/blog/what-is-git-bash

To jump start, here is a tutorial: Ubunto Linux for beginners.

At least, you need to know how to handle files and traverse across directories.

1.3.2 Python

Set up Python on your computer:

- Python 3.
- Python package manager **miniconda** or **pip**.
- Integrated Development Environment (IDE) (Jupyter Notebook; RStudio; VS Code; Emacs; etc.)

Readability is important! Check your Python coding styles against the recommended styles: https://peps.python.org/pep-0008/. A good place to start if the Section on "Code Lay-out".

1.4 Data Challenges

- ASA Data Challenge Expo
- Kaggle
- DrivenData
- Top 10 Data Science Competitions in 2023

1.5 Wishlist

This is a wish list from all members of the class (alphabetical order). Add yours; note the syntax of nested list in Markdown.

- Last, First
  - Complete ...
  - Become proficient ...
- Yan, Jun

- Make data science more accessible to undergraduates
- Co-develop a Quarto Book in collaboration with the students

1.5.1 Presentation Orders

Wait until next week to let the class size settle.

2 Project Management with Git

Many tutorials are available in different formats. Here is a YouTube video "Git and GitHub for Beginners — Crash Course'' The video also covers GitHub, a cloud service for Git. Other similar services are, for example, bitbucket and GitLab. A cloud service gives you a cloud back up of your work and makes collaboration with co-workers easy.

Here is a collection of online Git exersices that I used for Git training in other courses that I taught.

Tips on using Git:

- Use the command line interface instead of the web interface (e.g., upload on GitHub)
- Make frequent small commits instead of rare large commits.
- Make commit messages informative and meaningful.
- Name your files/folders by some reasonable convention.
  - Lower cases are better than upper cases.
  - No blanks in file/folder names.
- Keep the repo clean by not tracking generated files.
- Creat a `.gitignore` file for better output from `git status`.
- Keep the linewidth of sources to under 80 for better `git diff` view.

To set up GitHub (other services like Bitbucket or GitLab are similar), you need to

- Generate an SSH key if you don't have one already.
- Sign up an GitHub account.
- Add the SSH key to your GitHub account

The most frequently used Git commands are: + `git clone` + `git pull` + `git status` + `git add` + `git remove` + `git commit` + `git push`

3 Exercises

1. Practice Git.
    1. Clone the `ids-23` repo to your own computer.
    2. Add your name and wishes to the Wishlist; commit with an informative message.
    3. Remove the `Last, First` entry from the list; commit.
    4. Create a new file called `add.qmd` containing a few lines of texts; commit.
    5. Remove `add.qmd` (pretending that this is by accident; commit.
    6. Recover the accidently removed file `add.qmd`; commit.
    7. Put the repo into the GitHub Classroom homework repo and push.
    8. Make a pull-request to the `ids-23` repo.

References