Programming for Data Science

Rafael C. Alvarado

5/8/21

Table of contents

1 Preface

Welcome to Programming for Data Science, a collection of materials designed to support the course DS 5100 in the Data Science curriculum at UVA.

In this course, you will develop skills in Python and R Programming, as well as how to use the command line and GitHub.

The objective of this course is to introduce essential programming concepts, structures, and techniques.

You will gain confidence in not only reading code, but learning what it means to write good quality code.

Additionally, essential and complementary topics are taught, such as testing and debugging, exception handling, and an introduction to visualization.

Part I

Getting Started

During the technical orientation, you set up a GitHub account.

You also spent a little time browsing a sample repository, which you may wish to revisit:

https://github.com/UVADS/orientation-technical

You also should be able check off the following items:

- Understand the difference between Git and GitHub.
- Understand the purpose of Git and Github for data science work.
- Ensure Git is installed on your computer.
- Understand how to find a repository on GitHub.

2 Rivanna: UVA's HPC Cluster

2.1 Introduction

A useful infrastructural resource for this course is Rivanna, UVA's high-performance computing cluster. Each student has an account on Rivanna and access to resources there based on participation in this course.

Rivanna is UVA's High Performance Computing Cluster. We will use it in our class for both Python and R. This page describes some of the tools available for your use in this course. For information about Rivanna, see this introduction. Resources for getting help, including a knowledge base and ticket system, are found here www.rc.virginia.edu/support/.

You may need to use VPN to access Rivanna from an off-grounds location. To install VPN on your computer, go to the ITS VPN page for instructions. Note that you should connect to "UVA Anywhere," not to any of the higher security options. Course Allocation

This course has been allocated compute and space resources on Rivanna. The names of the resources are given below. The allocation ID needs to be entered to access certain tools. The storage path is accessible to you on the remote server.

- Allocation ID: `msds_ds5100`
- Storage path: `/project/MSDS_DS5100` ← don't use unless directed to

2.2 Tools

UVA Research Computing provides you with a suite of tools to access Rivanna. These tools are accessible through the menu on the UVA OpenOnDemand Dashboard page. Below are some brief descriptions of the tools.

File Explorer. A web-based GUI to access the file system of the remote server. Can be used to create, move, and delete directories and files, and to edit the contents of files (see Editor). You can also upload and download files through this interface. The File Explorer is useful to view your remote content and manage files and directories without having to use the command line. Note that not all operations can be performed through this interface.

Find under "Files" in the menu.

Editor. A web-based text editor launched from the File Explorer to view and edit text files on the remote server. Although not as sophisticated as VS Code (below), this is very useful for editing data and code files without having to use a command line editor. One advantage over VS Code is that it does not need to be launched – which means it does not time out like the Interactive Apps listed below.

The Editor is launched from the File Explorer.

SSH Shell Access (Terminal). Access to the command line of the remote server. Use this to open a terminal window to perform Linux commands directly. Note that It is necessary to use a terminal to install and run certain programs on the remote server.

Find under "Clusters" in the menu.

You can also access the remote command line via SSH on your local computer. Just enter the following on the command line of either a PC or Mac:

```
ssh -Y <userid>@hpc.rivanna.virginia.edu
```

Replace `<userid>` with your UVA Net ID, e.g. `abc2x`.

Be suer to be running VPN if you are accessing Rivanna from an off-grounds location.

2.3 Interactive Spps

These tools must be launched by specifying a set of parameters, including the allocation you are using. They are also timed and will close when time is up. Be sure to give yourself enough time when launching these, and to be aware of how much time you have when working.

Note also that you should allocate the fewest resources necessary to do the work you plan to do. This saves resources on the remote host, but also allows your app to launch more quickly. If you ask for an excessive amount of resources, you may wait a long time (e.g. hours) to have your app launched.

Desktop. Access to a GUI desktop to the remote server. This provides a access to various applications on the server, including a web browser, a file explorer, and terminal windows. Using this is not necessary if you can get by with the tools listed above.

Find under "Interactive Apps > Desktop" in the menu.

VS Code. Access to Visual Studio Code on the remote server. This is a fully functional instance of the IDE.

Find under "Interactive Apps > Code Server" in the menu.

Jupyter. Access to Jupyter Lab on the remote server. Find under "Interactive Apps > Jupyter Lab" in the menu.

RStudio. Access to Jupyter Lab on the remote server.

Find under "Interactive Apps > RStudio Server" in the menu.

2.4 For More Information

UVA's Research Computing unit provides resources for learning how to use Rivanna. Here are two slide decks that you may find useful:

- Introduction to Rivanna
- Using Rivanna from the Command Line

3 On Unix

3.1 Introduction

The Unix family of operating systems provide users with a command line interface (CLI) to execute commands and get things done. They also, typically provide GUIs but we won't go into those here.

The Unix family includes all varieties of Linux and the Mac OS (which is based on FreeBSD).

The command line that you actually interact with – the set of commands available to you – is called a shell, and there are several shells that you can run on your system. The most typical shell in use today is called bash which stands for Bourne Again Shell, since it is an improved version of bsh (The Bourne Shell). New versions of MacOS use the Z shell (zsh). The commands in these two shells are mostly similar, but there are subtle differences.

Windows has shells too for its command line interface. The default shell is DOS, but is also has PowerShell as an advanced (and very capable) option.

For more information, check out these resources:

- UVA Research Computing's Unix tutorial.
- Newham, 2005, Learning the bash Shell, O'Reilly Media.
- Jeroen Janssens, 2021, Data Science From the Command Line, O'Reilly Media.
- Neal Stephenson, 1999, In the Beginning Was The Command Line. (PDF version.)

3.2 Basic Commands

In this course, you don't need to know very many Unix shell commands, but you should be comfortable working from the command line to perform basic tasks. This is because some things can only be performed from the command line, such as installing some essential software. Here is a list of basic commands.

Navigating filesystems and managing directories:

- `cd` – change directory
- `pwd` – show the current directory
- `ln` – make links and symlinks to files and directories