

Open-Source Tools

Cheng Peng

West Chester University

This note introduces software programs and platforms that could be used in this course.

R & RStudio

1. What is R?

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R is an integrated suite of software facilities for data manipulation, calculation, and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple, and effective programming language which includes conditionals, loops, user-defined recursive functions, and input and output facilities.

– <https://www.r-project.org/about.html>

2. RStudio

RStudio is an integrated development environment (IDE) for R. It includes a console, and syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging, and workspace management.

There are two versions of RStudio: RStudio Desktop and RStudio Server. Both versions have free open-source and commercial editions. We use the free open-source edition of RStudio Desktop that has the following features:

- Access RStudio locally
- Syntax highlighting, code completion, and smart indentation
- Execute R code directly from the source editor
- Quickly jump to function definitions
- View content changes in real time with the Visual Markdown Editor
- Easily manage multiple working directories using projects
- Integrated R help and documentation

- Interactive debugger to diagnose and fix errors
- Extensive package development tools

3. The Relationship between R and RStudio

R and RStudio are two distinctly different applications that serve different purposes. R is a programming language used for statistical computing while RStudio uses the R language to develop statistical programs.

R and RStudio are not separate versions of the same program and cannot be substituted for one another. R may be used without RStudio, but RStudio may not be used without R.

RPubs

What is RPubs?

Register An Account with RPubs

First of all, you need to sign up for an account with RPubs if you don't have one. Otherwise, sign in to your existing RPubs account. The following two hyperlink buttons will bring you to the appropriate website.

Requirements

You need to install a recent version of R. You'll need R itself, RStudio, and the `knitr` package on your machine.

Steps for Publishing on RPubs

- In RStudio, create a new R Markdown document by choosing **File | New | R Markdown**.
- Click the **Knit HTML** button in the doc toolbar to preview your document.
- In the preview window, click button.

Github

What is Github?

GitHub is a social networking site for programmers to share their code. Many companies and organizations use it to facilitate project management and collaboration. It is the most prominent source code host, with over 60 million new repositories.

Most importantly, it is free. We can also use this resource to host web pages. Many images and data sets that I used are stored on GitHub.

Register A Github Account

You can use the following two buttons to sign up for an account with Github or sign in to an existing Github account.

Getting Started with GitHub

We will use screenshots to demonstrate how to create a repository, folders, and files.

1. After you logged into your account, you click the "continue for free" button located at the bottom of the following page (screenshot)
2. Now you see your Github front page. Click the green button "create repository" on the left panel. Our first repository is called "sta553"
3. To organize files in the repository `sta553`, We want folders for different files. To create a folder under `sta553`, click the hyperlink 'creating a new file'

4. The first folder to create is called the **data** folder which will be used to store data files. After typing “data/”, a new box appears under the “data” folder, type the first file name - readme, and the content of the file (see the screenshot). In the end, click the green button “Commit new file” to complete the creation of the first folder in the repository **data**.
5. To load the data file to the **data** folder, we click the drop-down menu on the top right corner and select ‘upload files
6. To create other folders under **sta553**, we click **Creating New File**, and we can create a new folder **image** similarly.
7. To create a new repository, Click the drop-down menu on the top right corner and select **New repository** to create a new repository.

SAS OnDemand

1. What is SAS OnDemand (SAS Studio)

SAS OnDemand provides **free** data management and data analysis tools. The advantage of SAS OnDemand is that it does not require any installation and it runs on the cloud via the internet and process data by connecting to the SAS server in the cloud. In other words, your computer is only used as a monitor since it does not use any resources (memory and CPU) of your computer.

Click **Access** to enter the SAS OnDemand login page.

2. Sign-in / Sign-up

If you have already created your SAS Profile, use the email or user ID and the password to log into the SAS OnDemand page.

3. Create An SAS Profile

If you don’t have a SAS profile, click the link **Don't have a SAS profile?**, and you will have the following pop-up dialogue box. Click **Create profile**, then you will see a pop-up sign-up page. You then follow the direction to create your SAS profile.

4. Log Into SAS Academic OnDemand

Provide your profile information to log into the OnDemand page, you will see the link to the SAS Studio user interface and your account information as well.

Once you created a SAS profile, you will have 5 GB of free storage.

5. SAS Studio User Interface

In the **Applications** tab, click **SAS Studio**, and you see the SAS Studio user interface on a separate page (it may take a little bit of time to initialize your account if you use it for the first time).

The above screenshot was taken from my SAS course webpage. For those who learned SAS using the classical SAS, you will see SAS Studio is much more convenient and easier to use.

A Cautionary Note on Data Security

SAS Studio (Academic OnDemand) is installed on SAS servers hosted in the Microsoft Azure Cloud. Although SAS claims that your assigned storage is private and secured, it is suggested to avoid uploading **sensitive data** to your private storage on the SAS server since SAS does not release the level of security for the storage.

R Viz Libraries

The following libraries will be used throughout this class.

1. Tidyverse

2. ggplot2

Ggplot2 is a system for creating charts based on the Grammar of Graphics. It proved to be one of the most powerful R libraries for visualization.

3. plotly

Plotly is an online platform for data visualization in R (also available in Python). This package creates interactive web-based plots using `plotly.js` library. Plotly gives users an opportunity to interact with graphs, change their scale and point out the necessary record. The library also supports graph hovering. Moreover, one can easily add Plotly in knitr/R Markdown or Shiny apps.

4. leaflet

Leaflet is a well-known package based on JavaScript libraries for interactive maps. It is widely used for mapping and working with the customization and design of interactive maps. Besides, Leaflet provides an opportunity to make these maps mobile-friendly.

5. mapview

6. tmap

7. Other infrequently used packages

ggmap, map, dygraph,