CSCI446/946 Big Data Analytics - Week 2

**Lab 1 - A brief introduction to Data Analytics using R and Python**

Table of Contents

[Introduction 2](#_Toc109996466)

[Task 1: Install R 3](#_Toc109996467)

[Task 2: Install Python 4](#_Toc109996468)

[Task 3: Install Anaconda 5](#_Toc109996469)

[Task 4: Install RStudio 6](#_Toc109996470)

[Task 5: Install Spyder 10](#_Toc109996471)

[Task 6: Exercise with R 13](#_Toc109996472)

[Task 7: Exercise with Python 14](#_Toc109996473)

[Task 8: Install Jupyter Notebook/Lab (Optional) 15](#_Toc109996474)

# Introduction

In Lab1, we will install and run R or/and Python on graphical editors. Graphical editors make coding, debugging, visualizing, results, etc., much more accessible. CSCI446/946 suggests using, but not limited to, **Rstudio** for R and **Spyder** for Python. **Jupyter Notebook** or **Juptyer Lab** are another good tools for CSCI446/946 labs and assignments (optional). **Anaconda** is a recommended application having RStudio, Spyder, Jupyter and managing packages of R and Python.

R and Python are often used in Big Data Analytics though more powerful tools exist (i.e., Hadoop). We use R and Python so that you can run lab tasks and exercises and do assignments on your home computer. All the tools we will use are available for a broad spectrum of OSs and are free to use!

You must choose **at least one** programming language (R or Python) for CSCI446/946 labs, exercises, and assignments. A combination use of R and Python is feasible. For example, in an assignment, Python is applied to run a clustering algorithm, and R is for evaluation and visualization.

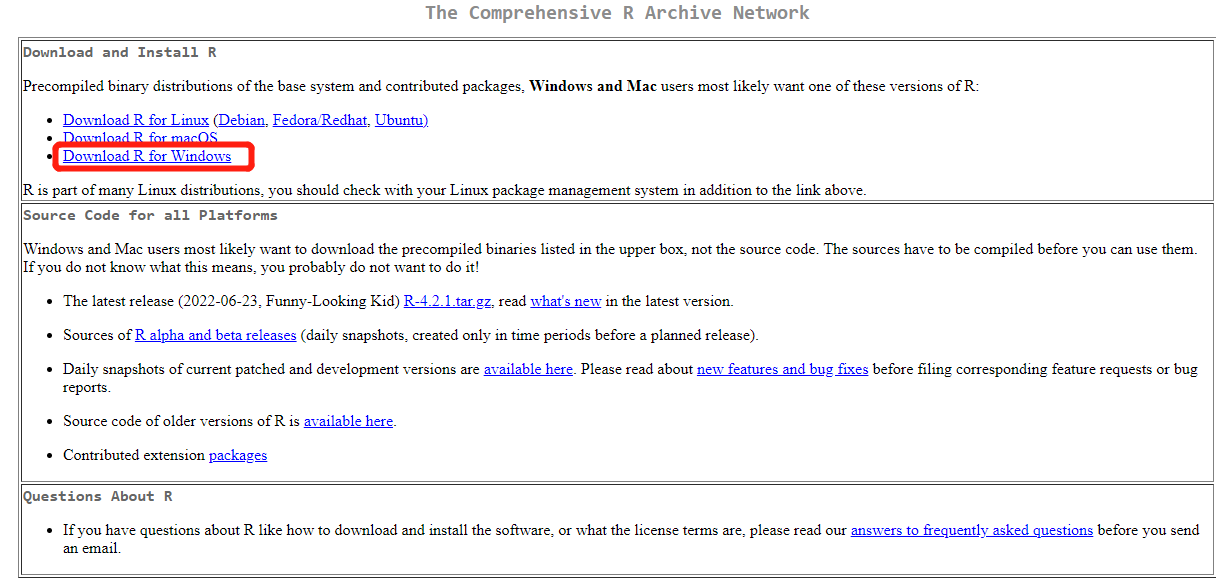
* If you and your assignment group choose to use R for programming, please complete **Task 1**, **Task3**, **Task4**, **Task6** below.
* If you and your assignment group choose to use Python for programming, please complete **Task 2**, **Task3**, **Task5**, **Task7**, **Task 8** (optional) below.
* If you and your assignment group choose to use both Python and R, please complete **all** Tasks below.

# Task 1: Install R

*(Move to the next task if you already have R 4.2.1 installed on your computer.)*

Step 1, To download R 4.1, visit: <https://cran.csiro.au>

Step 2, click on “Download R for Windows”



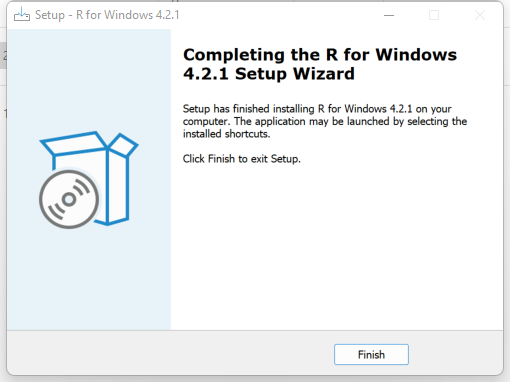
Step 3, click “base” or “install R for the first time”, then click “Download R-4.2.1 for Windows” to start download.

Graphical user interface, text, application

Description automatically generated



Step 3, find the downloaded file “R-4.2.1-win.exe” from your browser’s Downloads Location (default: Downloads folder) and run installation: multiple clicks on “Next”, then “Finish” to complete.



# Task 2: Install Python 3.10.5

*(Move to the next task if you already have Python 3.10.5 installed on your computer.)*

Step1, To download Python, visit: <https://www.python.org/downloads/> , click the button “Download Python 3.10.5” to download.

Graphical user interface, website

Description automatically generated

Step2, find the downloaded file “python-3.10.5-amd64.exe” from your browser’s Downloads Location (default: Downloads folder) and run installation: agree with default settings, then “Close” to finish.

Graphical user interface, text, application

Description automatically generated

# Task 3: Install Anaconda

*(Move to the next task if you already have Anaconda installed on your computer.)*

Step1, To download Anaconda, visit: <https://www.anaconda.com/> and click “Download” button.

Graphical user interface, text, application, email

Description automatically generated

Step2, find the downloaded file “Anaconda3-2022.05-Windows-x86\_64.exe” from your browser’s Downloads Location (default: Downloads folder) and run installation: multiple clicks on “Next” & “I Agree”, then “Finish” to complete.

Text

Description automatically generated

# Task 4: Install RStudio

*(Move to the next task if you already have RStudio installed on your computer.)*

Step 1, launch Anaconda (It is ok to update Anaconda to 2.2.0 if there is a pop-up window for updating), and install RStudio in latest version from Anaconda, click > install specific version > 1.1.456:

Graphical user interface, application

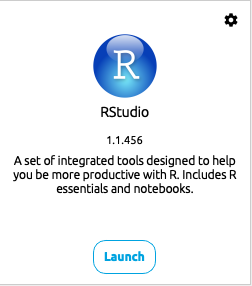
Description automatically generated

During the installation, there may be a pop-up window to create a new environment for RStudio. Use default environment name, i.e., “rstudio”, and create to continue.

Installation status bar is at the bottom of the window:



Once completed, we can launch RStudio from Anaconda every time by selecting “rstudio” from Anaconda > Environments first.



Step 2, download Rtools from <https://cran.csiro.au/> > “Download for Windows” > Rtools

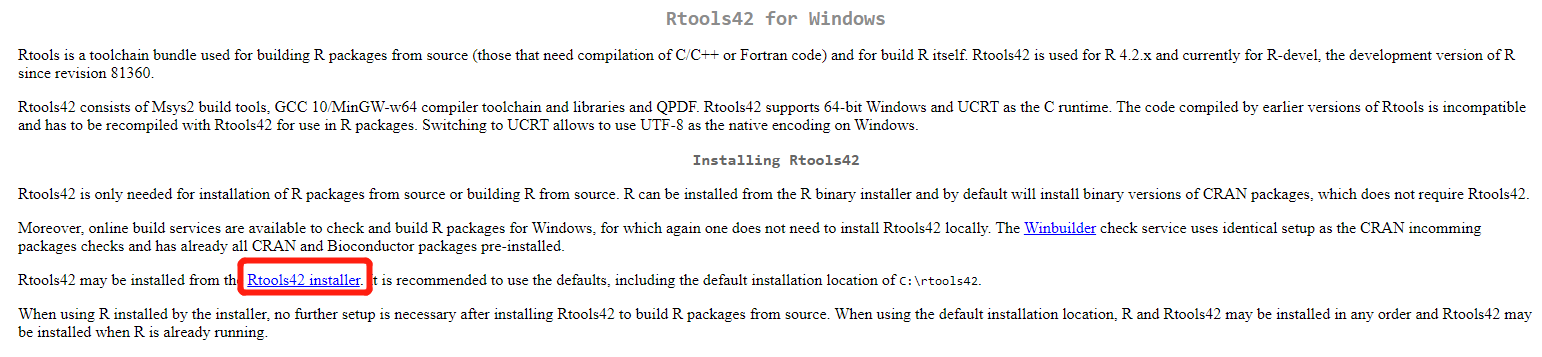
Text

Description automatically generated with medium confidence

Then, click “RTools-4.2” > click “Rtools42 installer” to download.

A picture containing text

Description automatically generated



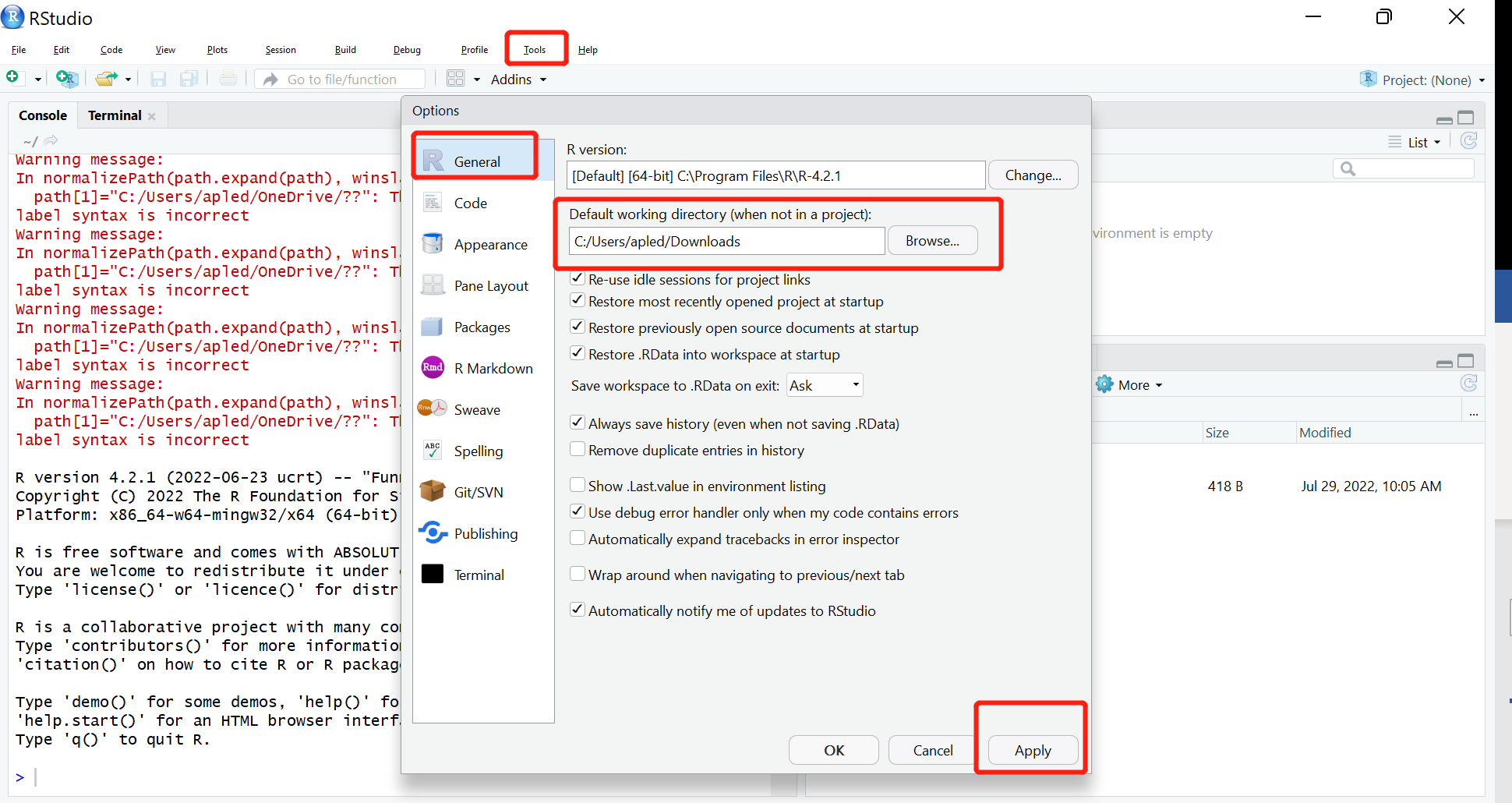
Find the downloaded file “rtools42-5253-5107-signed.exe” from Downloads and run installation: multiple clicks on “Next”, then “Install”, then “Finish” to complete.

Graphical user interface, application

Description automatically generated

Step3, configure RStudio. Firstly, launch RStudio from Anaconda (under rstudio environment). Then, in the menu-bar select “Tools”> “Global Options…”

Set the working directory to a directory where you wish to keep your R projects and associated files. I would change to Downloads.

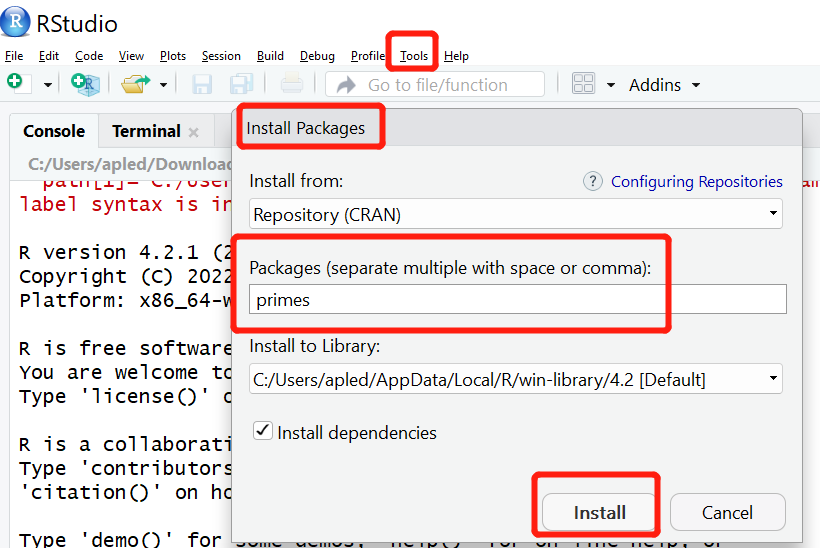


Then, install “primes” package to rstudio environment by Anaconda.

Graphical user interface, application, Word

Description automatically generated

Restart RStuodio from Anaconda under rstudio environment. Select “Tools” > “Install Packages”, then install “primes” packages.



Step4, test RStudio & familiarize yourself with R.

1. Start Rstudio
2. In the console, type:

library(primes)

generate\_primes(min=0,max=100)



*If you are new to R then follow online tutorials such as* [*https://www.w3schools.com/r/*](https://www.w3schools.com/r/)

Ref: <https://cran.r-project.org/web/packages/primes/primes.pdf>

# Task 5: Install Spyder

*(Move to the next task if you already have Spyder installed on your computer.)*

Step 1, launch Anaconda

Step 2, go to Environments, then create a new environment for python

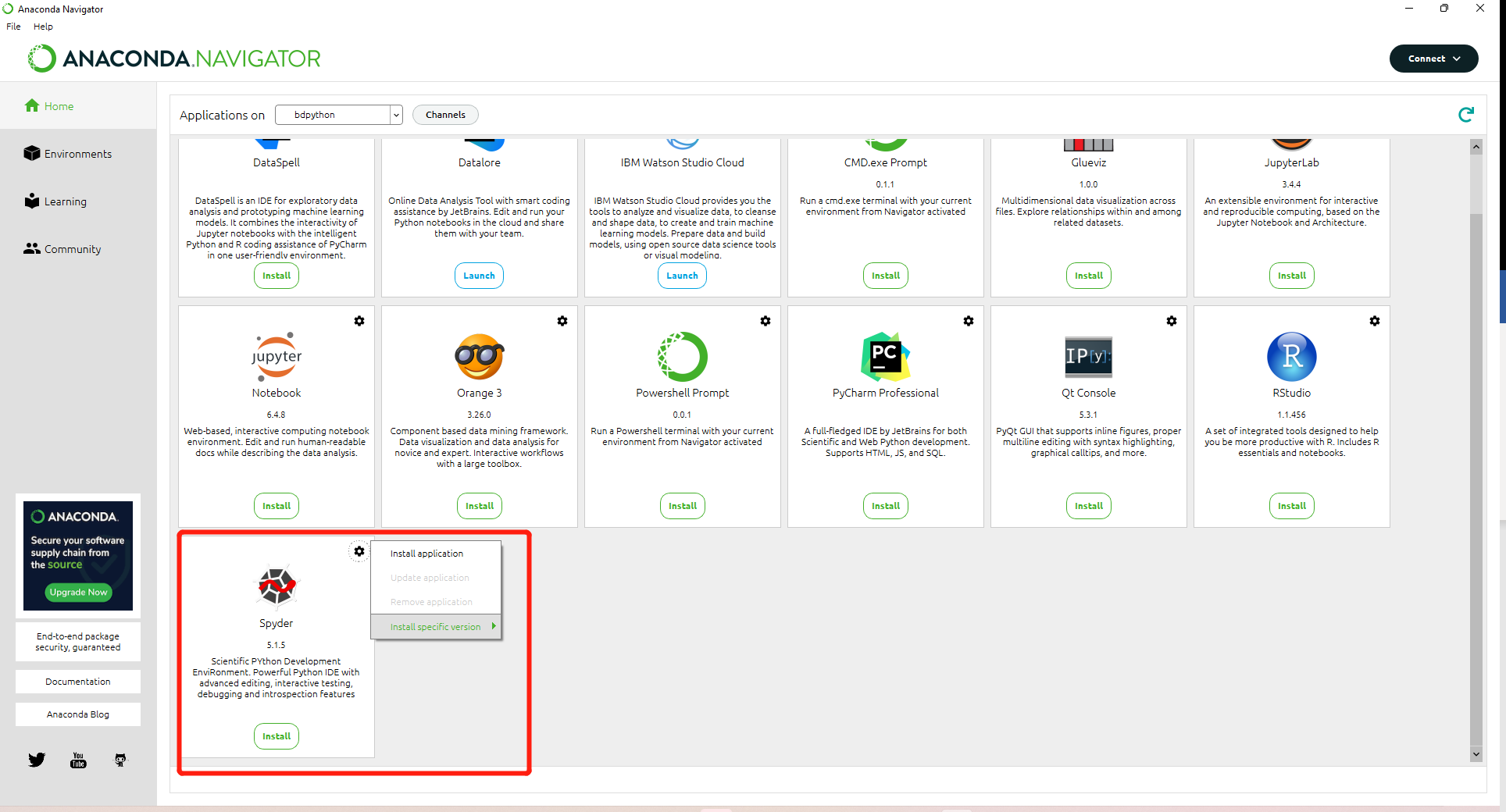
Graphical user interface, application

Description automatically generated

Installation status bar is at the bottom of the window:



Step3, by selecting bdpython in Environments, go to Home, then install Spyder in latest version from Anaconda, click> install specific version > 5.1.5 :

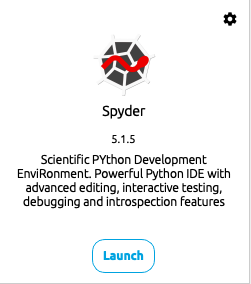


During the installation, there may be a pop-up window to create a new environment for spyder. Use default environment name, i.e., “spyder”, and create to continue.

Installation status bar is at the bottom of the window:

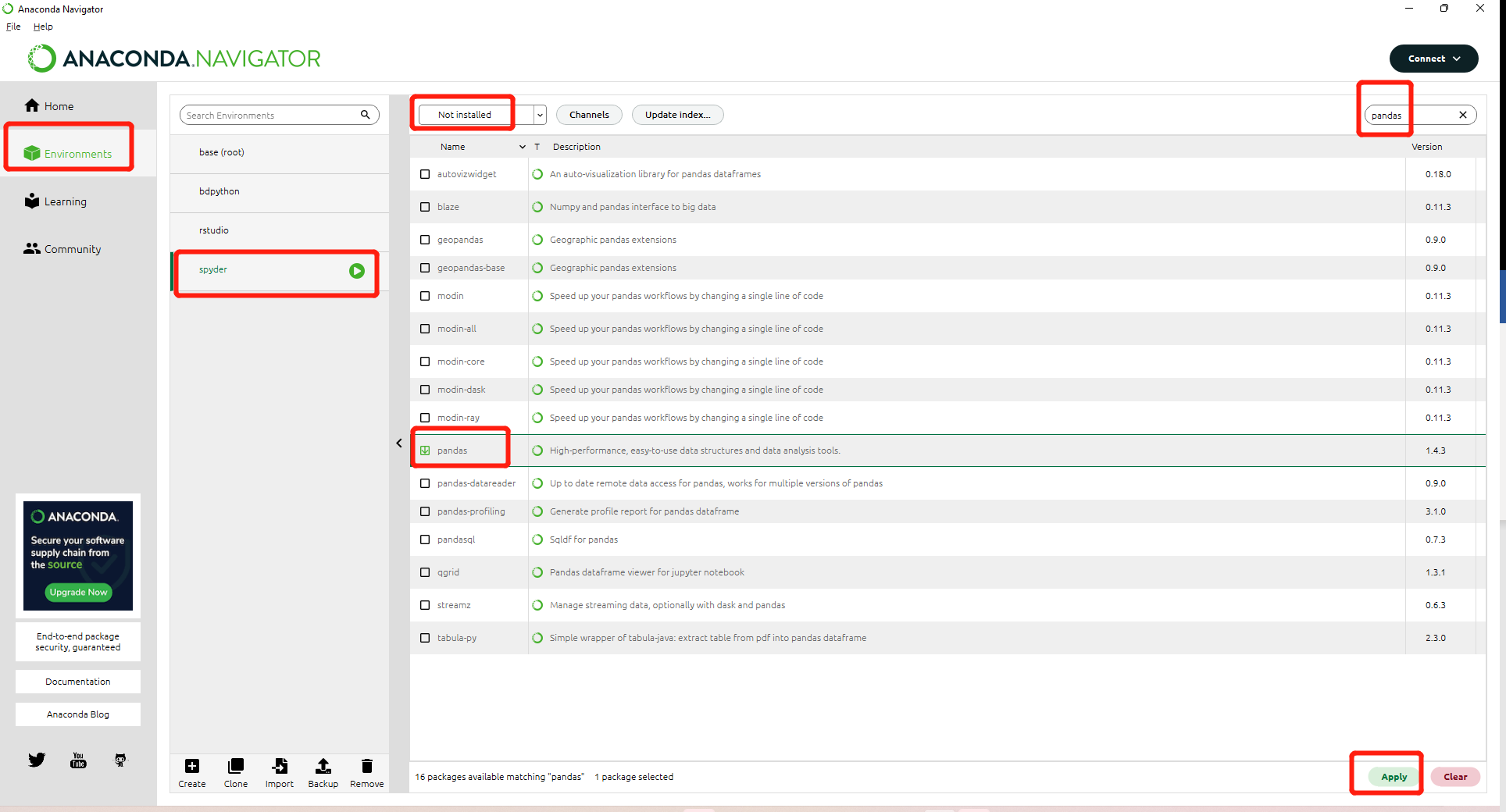


Once completed, we can launch Spyder from Anaconda every time by selecting the spyder environment first from Anaconda > Environments.



Step4, install Python packages. There are many packages and toolboxes available for Python. We can install a particular package when we need it. There are two ways to install packages: Anaconda -> Environments, or cmd (Windows command console) terminal command. Take the package pandas for example:

1. Anaconda > Environments > spyder



1. Terminal command:

Firstly, From Anaconda to open environment spyder in terminal: Environments > spyder >  > Open Terminal

Graphical user interface, text, application, chat or text message

Description automatically generated

Then, type pip install pandas in terminal

Text

Description automatically generated

Use either way to install 3 packages: pandas, numpy, sklearn into spyder environment.

Step 5, change default dictionary to Downloads folder in Spyder: menu-bar > Tools > PYTHONPATH manager > Add path > select Downloads folder > click Select Folder > OK, then menu-bar > Consoles > Restart kernel, then menu-bar > File > New file… .

A screenshot of a computer

Description automatically generated with medium confidence

# Task 6: Exercise with R

Let’s do some exercised with R to gain some familiarity.

**[Exercise]** Run this example with RStudio. You can copy-paste those examples but you

should understand what each function does. The .csv file can be downloaded from Moodle under Week2 section. *With RStudio default directory under Downloads, .csv file should be kept in Downloads.*

# read a csv file of the total annual sales for each customer

sales <- read.csv("./yearly\_sales.csv")

# examine the imported dataset

head(sales)

summary(sales)

# plot num\_of\_orders vs. sales

plot(sales$num\_of\_orders,sales$sales\_total, main="Number of Orders vs. Sales")

# perform a statistical analysis (fit a linear regression model)

results <- lm(sales$sales\_total ~ sales$num\_of\_orders)

summary(results)

# perform some diagnostics on the fitted model

# plot histogram of the residuals

hist(results$residuals, breaks = 800)

# add a column for the average sales per order

sales$per\_order <- sales$sales\_total/sales$num\_of\_orders

# export data as tab delimited without the row names

write.table(sales,"sales\_modified.txt", sep="\t", row.names=FALSE)

# export a histogram to a jpeg

jpeg(file="sales\_hist.jpeg") # create a new jpeg

file

hist(sales$num\_of\_orders) # export histogram to jpeg

dev.off() # shut off the graphic device

# Task 7: Exercise with Python

Let’s do some exercised with Python to gain some familiarity.

**[Exercise]** Launch Spyder then run each of the following commands. Notice that the following does exactly the same as what we did in R in Task 6.

*For preparation, install package matplotlib first.*

import pandas as pd

import matplotlib.pyplot as plt

# read the csv file

sales = pd.read\_csv('yearly\_sales.csv')

# examine the imported dataset

sales.head()

sales.describe()

# plot num\_of\_orders vs. sales

sales.plot(x='num\_of\_orders', y='sales\_total', style='o')

# Task 8: Install Jupyter Notebook/Lab (Optional)

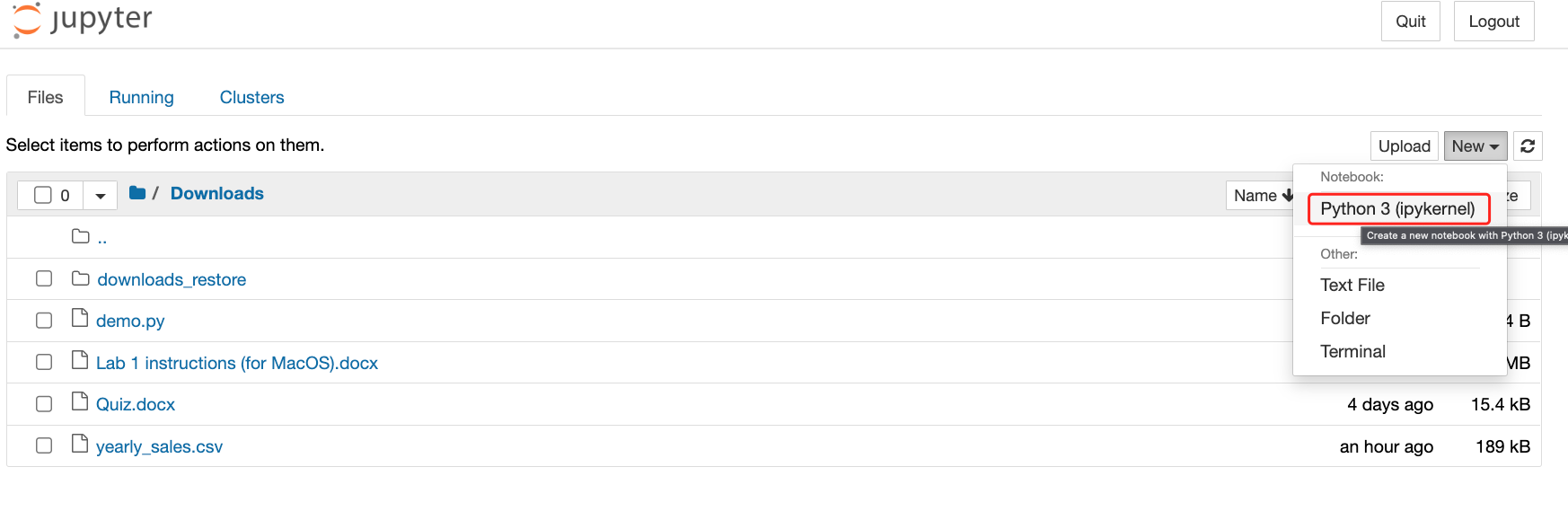
Step 1, launch Anaconda

Step 2, create a new environment or use spyder or bdpython from Environments. (Following use spyder)

Step 3, select spyder, then go to Home, then click Install under Jupter Notebook.

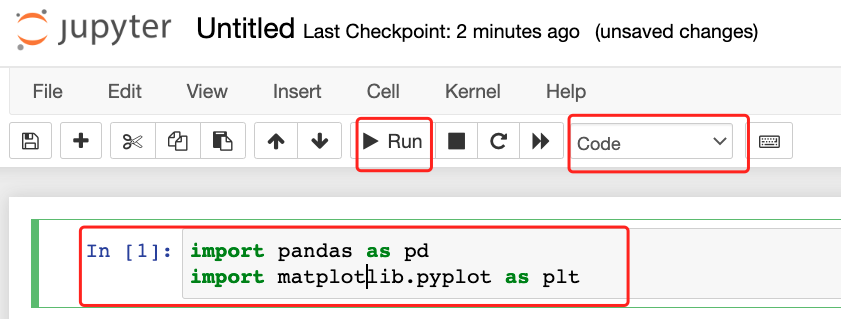
Step 4, launch Jupter Notebook from Anaconda. Jupter Notebook is opened in a browser.

Step 5, Select Downloads folder, then create a new notebook.



In the jupyter notebook,  run a cell a time and create a new cell after running the previous cell.

It uses a Code cell to run python command.

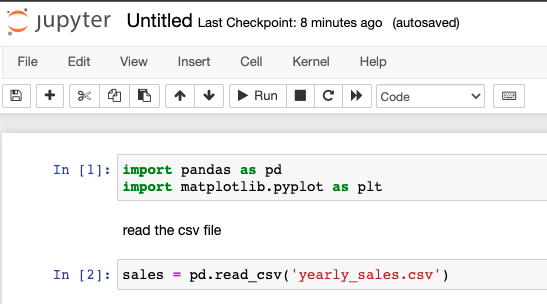


It uses Markdown cell to display plain words.

Graphical user interface, text, application

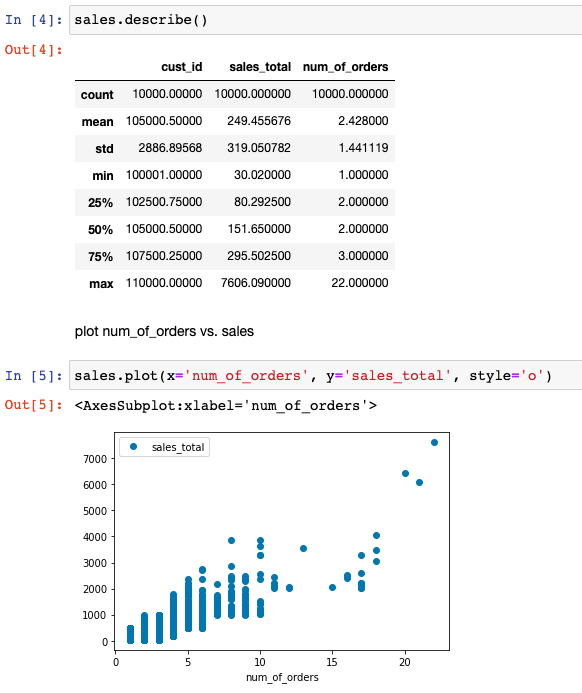
Description automatically generated

Please apply codes in Task 7 in a jupyter notebook. The formatted notebook is illustrated as below.

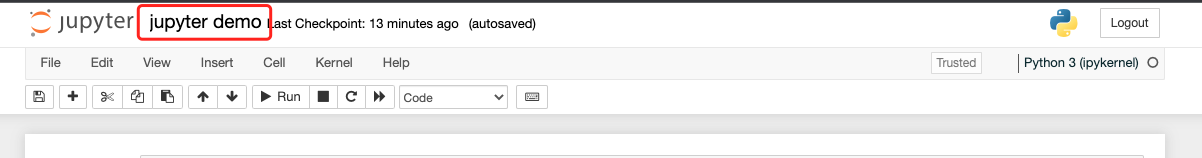


Table

Description automatically generated



Single click to rename the jupyter notebook. The file is saved under Downloads.



More info in <https://jupyter.org/>