Python/R for Data Science 2022 Fall

Xinli Xiao

2022-08-01T14:43:17-05:00

Table of contents

Preface

This is a Quarto book.

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

Part I

Part I: Python

1 Python Fundamentals

1.1 Why Python?

1.1.1 Python is easy to use

Programmers familiar with traditional languages will find it easy to learn Python. All of the familiar constructs—loops, conditional statements, arrays, and so forth—are included, but many are easier to use in Python. Here are a few of the reasons why: - Types are associated with objects, not variables. A variable can be assigned a value of any type, and a list can contain objects of many types. This also means that type casting usually isn't necessary and that your code isn't locked into the straitjacket of predeclared types. - Python typically operates at a much higher level of abstraction. This is partly the result of the way the language is built and partly the result of an extensive standard code library that comes with the Python distribution. A program to download a web page can be written in two or three lines! - Syntax rules are very simple. Although becoming an expert Pythonista takes time and effort, even beginners can absorb enough Python syntax to write useful code quickly.

Python is well suited for rapid application development. It isn't unusual for coding an application in Python to take one-fifth the time it would in C or Java and to take as little as one-fifth the number of lines of the equivalent C program. This depends on the particular application, of course; for a numerical algorithm performing mostly integer arithmetic in for loops, there would be much less of a productivity gain. For the average application, the productivity gain can be significant.

1.1.2 Python is expressive

1.1.3 Python is readable

[1] [2] [3] [4] [5] [6]

1.2 Hello world!

We will mainly focus on this code editor mode at the beginning and check our results or do some simple computations in the console.

Notebook is another very popular mode to use Python. We will talk about it later.

1.2.1 Set up Python environment using IDEs

Please follow the following steps to run your first line of Python codes.

We will talk about the relation between Python and Anaconda and more about packages sometime later.

1. Go to Anaconda download page. Download and install Anaconda.
2. There are several IDEs for Python bundled with Anaconda. Pick any one you like. I personally use VS Code. Here we use Spyder as an example for now since it doesn't require any configurations.
3. Here is a screenshot of Spyder 5.1.5.
4. The right lower window is the console. Type the following code, and run. If `Hello world!` is displayed, the Python environment is set up successfully. Now you can start to play with Python!

1.2.2 Code editor

The left window is called *Code Editor*. You can write multiple lines of codes in the code editor and run them all together. The output results might appear in the console.

As shown in the screenshot, when press `F5` to run file, the codes in the code editor will be excuted line by line.

The code in the example is

```
print('Hello world!')
```

```
Hello world!
```

```
print('Another line')
# Everything after # are comments that won't be excuted.
```

```
Another line
```

1.2.3 Indentation

One key feature about Python is that its structures (blocks) is determined by **Indentation**.

Let's compare with other languages. Let's take `C` as an example.

```
#/*This is a C function.*/
#int f(int x){return x;}
```

The block is defined by `{}` and lines are separated by `;`. `space` and `newline` are not important when `C` runs the code. It is recommended to write codes in a "beautiful, stylish" format for readibility, as follows. However it is not mandatary.

```
#/*This is a C function.*/
#int f(int x) {
#    return x;
#}
```

In Python, blocks starts from `:` and then are determined by indents. Therefore you won't see a lot of `{}` in Python, and the "beautiful, stylish" format is mandatary.

```
# This is a Python function.
def f(x):
    return x
```

The default value for indentation is 4 spaces, which can be changed by users. We will just use the default value in this course.

1.3 Built-in Types

There are severl built-in data structures in Python. Here is an (incomplete) list: - `None` - Boolean – `True`, `False` - Numeric Types — `int`, `float`, `complex` - Sequence Types — `list` - Text Sequence Type — `str` - Map type - `dict`

We will cover numeric types and strings in this section. The rests are either simple that are self-explained, or not simple that will be discussed later.

You may always use `type(x)` to detect the type of the object `x`.

1.3.1 Numeric types and math expressions

Numeric types are represented by numbers. If there are no confusions, Python will automatically detect the type.

```
x = 1 # x is an int.
y = 2.0 # y is a float.
```

Python can do math just like other programming languages. The basic math operations are listed as follows. - `+`, `-`, `*`, `/`, `>`, `<`, `>=`, `<=` works as normal. - `**` is the power operation. - `%` is the mod operation. - `!=` is `not equal`

1.3.2 Strings

Scalars are represented by numbers and strings are represented by quotes. Example:

```
x = 1       # x is a scalar.
y = 's'     # y is a string with one letter.
z = '0'     # z loos like a number, but it is a string.
w = "Hello" # w is a string with double quotes.
```

Here are some facts. 1. For strings, you can use either single quotes or double quotes. 2. `\` is used to denote escaped words. You may find the list Here. 3. There are several types of scalars, like `int`, `float`, etc.. Usually Python will automatically determine the type of the data, but sometimes you may still want to declare them manually. 4.

Although `str` is a built-in type, there are tons of tricks with `str`, and there are tons of packages related to strings. Generally speaking, to play with strings, we are interested in two types of questions. - Put information together to form a string. - Extract information from a string. We briefly talk about these two tasks.

1.3.2.1 concat

1.3.2.2 split

There is a very subtle relations between the variable / constant and the name of the variable / constant. We

1.4 Flows

1.5 Logic

1.6 Exercises

Exercise 1.1 (Indentation)**.** Please tell the differences between the following codes. If you don't understand `for` don't worry about it. Just focus on the indentation and try to understand how the codes work.

```
for i in range(5):
    print('Hello world!')
print('Hello world!')
```

```
for i in range(5):
    print('Hello world!')
    print('Hello world!')
```

```
for i in range(5):
print('Hello world!')
print('Hello world!')
```

```
for i in range(5):
    pass
print('Hello world!')
print('Hello world!')
```

Exercise 1.2 (Play with built-in data types)**.**

```
print(True and True)
```

```
True
```

```
print(True or True)
```

```
True
```

```
print(False and True)
```

```
False
```

```
print((1+1>2) or (1-1<1))
```

```
True
```

Exercise 1.3 (Play with strings)**.**

1.7 Projects

jupyter notebook

1.7.1 Gnomic data

lists of data

2 more advanced techniques

2.1 List

append

extend

2.1.1 List Comprehension

List Comprehension is a convenient way to create lists based on the values of an existing list. It cannot provide any real improvement to the performance of the codes, but it can make the codes shorter and easier to read.

The format of list Comprehension is

```
newlist = [expression for item in iterable if condition == True]
```

> hehehe

2.2 dict

2.3 numpy and pandas

2.4 exercises

2.5 proj

2.5.1 key

2.5.2 project

- open and parse a CSV file

- do some operations
- decipher (replace letters / frequency analysis)

3 Classes/Packages for Python

3.1 functions

3.2 packages

3.3 classes

3.4 virenev

3.5 ex

3.6 proj

- clean commented code, use of functions
- write classes
- folllow-up: geospatial data based analysis in python

4 Python Notebooks

4.1 notebooks

4.2 ex

4.3 proj

Part II

Part II: R

5 R Fundamentals

[7]

A few advantages about R:

- Free and open source comparing to some other tools like Excel and SPSS.
- Optimized with vectorization.

5.1 Hello world for R

References

[1] Klosterman, S. (2021). *Data science projects with python: A case study approach to gaining valuable insights from real data with machine learning*. Packt Publishing, Limited.

[2] Ramalho, L. (2015). *Fluent python : Clear, concise, and effective programming: Clear, concise, and effective programming*. O'Reilly Media.

[3] Shaw, Z. A. (2017). *Learn python 3 the hard way*. Addison Wesley.

[4] Shaw, Z. A. (2017). *Learn more python 3 the hard way: The next step for new python programmers*. ADDISON WESLEY PUB CO INC.

[5] Ceder, N. R. (2018). *The quick python book*. MANNING PUBN.

[6] Youens-Clark, K. (2020). *Tiny python projects*. Manning Publications.

[7] Wickham, H. and Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.