

Introduction to programming for data science

STAT 201

Arvind Krishna

9/20/2022

Table of contents

Preface	3
1 Introduction to Jupyter Notebooks and programming in python	4
1.1 Jupyter notebook	4
1.1.1 Introduction	4
1.1.2 Writing and executing code	5
1.1.3 Saving and loading notebooks	5
1.1.4 Rendering notebook as HTML	6
1.2 In-class exercise	6
1.3 Debugging and errors	6
1.4 Terms used in programming	6
2 Data structures	7
2.0.1 Concatenating tuples	8
2.0.2 Unpacking tuples	8
2.0.3 Tuple methods	9
2.1 List	10
2.1.1 Adding and removing elements in a list	10
2.1.2 Concatenating lists	12
2.1.3 Sorting a list	13
2.1.4 Slicing a list	13
2.2 Dictionary	16
2.2.1 Adding and removing elements in a dictionary	16
2.3 Functions	18
2.3.1 Global and local variables with respect to a function	19
References	20

Preface

This book is currently being written for the course STAT201.

1 Introduction to Jupyter Notebooks and programming in python

This chapter is a very brief introduction to python and Jupyter notebooks. We only discuss the content relevant for applying python to analyze data.

Anaconda: If you are new to python, we recommend downloading the [Anaconda installer](#) and following the instructions for installation. Once installed, we'll use the Jupyter Notebook interface to write code.

Quarto: We'll use Quarto to publish the `**ipynb*` file containing text, python code, and the output. Download and install Quarto from [here](#).

1.1 Jupyter notebook

1.1.1 Introduction

Jupyter notebook is an interactive platform, where you can write code and text, and make visualizations. You can access Jupyter notebook from the Anaconda Navigator, or directly open the Jupyter Notebook application itself. It should automatically open up in your default browser. The figure below shows a Jupyter Notebook opened with Google Chrome. This page is called the *landing page* of the notebook.

<IPython.core.display.Image object>

To create a new notebook, click on the **New** button and select the **Python 3** option. You should see a blank notebook as in the figure below.

<IPython.core.display.Image object>

1.1.2 Writing and executing code

Code cell: By default, a cell is of type *Code*, i.e., for typing code, as seen as the default choice in the dropdown menu below the *Widgets* tab. Try typing a line of python code (say, `2+3`) in an empty code cell and execute it by pressing *Shift+Enter*. This should execute the code, and create an new code cell. Pressing *Ctrl+Enter* for *Windows* (or *Cmd+Enter* for *Mac*) will execute the code without creating a new cell.

Commenting code in a code cell: Comments should be made while writing the code to explain the purpose of the code or a brief explanation of the tasks being performed by the code. A comment can be added in a code cell by preceding it with a `#` sign. For example, see the comment in the code below.

Writing comments will help other users understand your code. It is also useful for the coder to keep track of the tasks being performed by their code.

```
#This code adds 3 and 5
3+5
```

8

Markdown cell: Although a comment can be written in a code cell, a code cell cannot be used for writing headings/sub-headings, and is not appropriate for writing lengthy chunks of text. In such cases, change the cell type to *Markdown* from the dropdown menu below the *Widgets* tab. Use any markdown cheat sheet found online, for example, [this one](#) to format text in the markdown cells.

Give a name to the notebook by clicking on the text, which says ‘Untitled’.

1.1.3 Saving and loading notebooks

Save the notebook by clicking on **File**, and selecting **Save as**, or clicking on the **Save and Checkpoint** icon (below the **File** tab). Your notebook will be saved as a file with an extension *ipynb*. This file will contain all the code as well as the outputs, and can be loaded and edited by a Jupyter user. To load an existing Jupyter notebook, navigate to the folder of the notebook on the *landing page*, and then click on the file to open it.

1.1.4 Rendering notebook as HTML

We'll use Quarto to print the `**ipynb*` file as HTML. Check the procedure for rendering a notebook as HTML [here](#). You have several options to format the file.

You will need to open the command prompt, navigate to the directory containing the file, and use the command: `quarto render filename.ipynb --to html`.

1.2 In-class exercise

1. Create a new notebook.
2. Save the file as `In_class_exercise_1`.
3. Give a heading to the file - `First HTML file`.
4. Print `Today is day 1 of my programming course`.
5. Compute and print the number of seconds in a day.

The HTML file should look like the picture below.

`<IPython.core.display.Image object>`

1.3 Debugging and errors

Read sections 1.3 - 1.6 from http://openbookproject.net/thinkcs/python/english3e/way_of_the_program.html

1.4 Terms used in programming

Read section 1.11 from http://openbookproject.net/thinkcs/python/english3e/way_of_the_program.html

2 Data structures

In this chapter we'll learn about the python data structures that are often used or appear while analyzing data.

Tuple is a sequence of python objects, with two key characteristics: (1) the number of objects are fixed, and (2) the objects are immutable, i.e., they cannot be changed.

Tuple can be defined as a sequence of python objects separated by commas, and enclosed in rounded brackets (). For example, below is a tuple containing three integers.

```
tuple_example = (2,7,4)
```

We can check the data type of a python object using the *type()* function. Let us check the data type of the object *tuple_example*.

```
type(tuple_example)
```

tuple

Elements of a tuple can be extracted using their index within square brackets. For example the second element of the tuple *tuple_example* can be extracted as follows:

```
tuple_example[1]
```

7

Note that an object of a tuple cannot be modified. For example, consider the following attempt in changing the second element of the tuple *tuple_example*.

```
tuple_example[1] = 8
```

TypeError: 'tuple' object does not support item assignment

The above code results in an error as tuple elements cannot be modified.

2.0.1 Concatenating tuples

Tuples can be concatenated using the + operator to produce a longer tuple:

```
(2,7,4) + ("another", "tuple") + ("mixed","datatypes",5)
```

```
(2, 7, 4, 'another', 'tuple', 'mixed', 'datatypes', 5)
```

Multiplying a tuple by an integer results in repetition of the tuple:

```
(2,7,"hi") * 3
```

```
(2, 7, 'hi', 2, 7, 'hi', 2, 7, 'hi')
```

2.0.2 Unpacking tuples

If tuples are assigned to an expression containing multiple variables, the tuple will be unpacked and each variable will be assigned a value as per the order in which it appears. See the example below.

```
x,y,z = (4.5, "this is a string", ("Nested tuple",5))
```

```
x
```

```
4.5
```

```
y
```

```
'this is a string'
```

```
z
```

```
('Nested tuple', 5)
```

If we are interested in retrieving only some values of the tuple, the expression `*_` can be used to discard the other values. Let's say we are interested in retrieving only the first and the last two values of the tuple:


```
x,*_,y,z = (4.5, "this is a string", (("Nested tuple",5)), "99",99)
```

```
x
```

```
4.5
```

```
y
```

```
'99'
```

```
z
```

```
99
```

2.0.3 Tuple methods

A couple of useful tuple methods are `count`, which counts the occurrences of an element in the tuple and `index`, which returns the position of the first occurrence of an element in the tuple:

```
tuple_example = (2,5,64,7,2,2)
```

```
tuple_example.count(2)
```

```
3
```

```
tuple_example.index(2)
```

```
0
```

Now that we have an idea about tuple, let us try to think where it can be used.

```
<IPython.core.display.HTML object>
```

2.1 List

List is a sequence of python objects, with two key characteristics that differentiate them from tuples: (1) the number of objects are variable, i.e., objects can be added or removed from a list, and (2) the objects are mutable, i.e., they can be changed.

List can be defined as a sequence of python objects separated by commas, and enclosed in square brackets []. For example, below is a list containing three integers.

```
list_example = [2,7,4]
```

2.1.1 Adding and removing elements in a list

We can add elements at the end of the list using the *append* method. For example, we append the string 'red' to the list *list_example* below.

```
list_example.append('red')
```

```
list_example
```

```
[2, 7, 4, 'red']
```

Note that the objects of a list or tuple can be of different datatypes.

An element can be added at a specific location of the list using the *insert* method. For example, if we wish to insert the number 2.32 as the second element of the list *list_example*, we can do it as follows:

```
list_example.insert(1,2.32)
```

```
list_example
```

```
[2, 2.32, 7, 4, 'red']
```

For removing an element from the list, the *pop* and *remove* methods may be used. The *pop* method removes an element at particular index, while the *remove* method removes the element's first occurrence in the list by its value. See the examples below.

Let us say, we need to remove the third element of the list.

```
list_example.pop(2)
```

7

```
list_example
```

```
[2, 2.32, 4, 'red']
```

Let us say, we need to remove the element 'red'.

```
list_example.remove('red')
```

```
list_example
```

```
[2, 2.32, 4]
```

```
#If there are multiple occurrences of an element in the list, the first occurrence will be removed
list_example2 = [2,3,2,4,4]
list_example2.remove(2)
list_example2
```

```
[3, 2, 4, 4]
```

For removing multiple elements in a list, either `pop` or `remove` can be used in a for loop, or a for loop can be used with a condition. See the examples below.

Let's say we need to remove integers less than 100 from the following list.

```
list_example3 = list(range(95,106))
list_example3
```

```
[95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105]
```

```
#Method 1: For loop with remove
list_example3_filtered = list(list_example3) #
```

```

for element in list_example3:
    #print(element)
    if element<100:
        list_example3_filtered.remove(element)
print(list_example3_filtered)

```

[100, 101, 102, 103, 104, 105]

Q: What's the need to define a new variable 'list_example3_filtered' in the above code?

Replace list_example3_filtered with list_example3 and identify the issue.

```

#Method 2: For loop with condition
[element for element in list_example3 if element>100]

```

[101, 102, 103, 104, 105]

2.1.2 Concatenating lists

As in tuples, lists can be concatenated using the + operator:

```

import time as tm

list_example4 = [5,'hi',4]
list_example4 = list_example4 + [None,'7',9]
list_example4

```

[5, 'hi', 4, None, '7', 9]

For adding elements to a list, the **extend** method is preferred over the + operator. This is because using the + operator creates a new list, while the **extend** method adds elements to an existing list.

```

list_example4 = [5,'hi',4]
list_example4.extend([None, '7', 9])
list_example4

```

[5, 'hi', 4, None, '7', 9]

2.1.3 Sorting a list

A list can be sorted using the `sort` method:

```
list_example5 = [6,78,9]
list_example5.sort(reverse=True) #the reverse argument is used to specify if the sorting i
list_example5
```

[78, 9, 6]

2.1.4 Slicing a list

We may extract or update a section of the list by passing the starting index (say `start`) and the stopping index (say `stop`) as `start:stop` to the index operator `[]`. This is called *slicing* a list. For example, see the following example.

```
list_example6 = [4,7,3,5,7,1,5,87,5]
```

Let us extract a slice containing all the elements starting from the the 3rd position upto the 7th position.

```
list_example6[2:7]
```

[3, 5, 7, 1, 5]

Note that while the element at the `start` index is included, the element with the `stop` index is excluded in the above slice.

If either the `start` or `stop` index is not mentioned, the slicing will be done from the beginning or upto the end of the list, respectively.

```
list_example6[:7]
```

[4, 7, 3, 5, 7, 1, 5]

```
list_example6[2:]
```

[3, 5, 7, 1, 5, 87, 5]

To slice the list relative to the end, we can use negative indices:

```
list_example6[-4:]
```

```
[1, 5, 87, 5]
```

```
list_example6[-4:-2:]
```

```
[1, 5]
```

An extra colon (':') can be used to slice every ith element of a list.

```
#Selecting every 3rd element of a list  
list_example6[::3]
```

```
[4, 5, 5]
```

```
#Selecting every 3rd element of a list from the end  
list_example6[::-3]
```

```
[5, 1, 3]
```

```
#Selecting every element of a list from the end or reversing a list  
list_example6[::-1]
```

```
[5, 87, 5, 1, 7, 5, 3, 7, 4]
```

Now that we have an idea about lists, let us try to think where it can be used.

<IPython.core.display.HTML object>

Now that we have learned about lists and tuples, let us compare them.

Q: A list seems to be much more flexible than tuple, and can replace a tuple almost everywhere. Then why use a tuple?

A: The additional flexibility of a list comes at the cost of efficiency. Some of the advantages of a tuple over a list are as follows:

1. Since a list can be extended, space is over-allocated when creating a list. A tuple takes less storage space as compared to a list of the same length.
2. Tuples are not copied. If a tuple is assigned to another tuple, both tuples point to the same memory location. However, if a list is assigned to another list, a new list is created consuming the same memory space as the original list.
3. Tuples refer to their element directly, while in a list, there is an extra layer of pointers that refers to their elements. Thus it is faster to retrieve elements from a tuple.

The examples below illustrate the above advantages of a tuple.

```
#Example showing tuples take less storage space than lists for the same elements
tuple_ex = (1, 2, 'Obama')
list_ex = [1, 2, 'Obama']
print("Space taken by tuple =",tuple_ex.__sizeof__()," bytes")
print("Space taken by list =",list_ex.__sizeof__()," bytes")
```

Space taken by tuple = 48 bytes
Space taken by list = 64 bytes

```
#Examples showing that a tuples are not copied, while lists can be copied
tuple_copy = tuple(tuple_ex)
print("Is tuple_copy same as tuple_ex?", tuple_ex is tuple_copy)
list_copy = list(list_ex)
print("Is list_copy same as list_ex?",list_ex is list_copy)
```

Is tuple_copy same as tuple_ex? True
Is list_copy same as list_ex? False

```
#Examples showing tuples takes lesser time to retrieve elements
import time as tm
tt = tm.time()
list_ex = list(range(1000000)) #List containinig whole numbers upto 1 million
a=(list_ex[::-2])
print("Time take to retrieve every 2nd element from a list = ", tm.time()-tt)

tt = tm.time()
tuple_ex = tuple(range(1000000)) #tuple containinig whole numbers upto 1 million
a=(tuple_ex[::-2])
print("Time take to retrieve every 2nd element from a tuple = ", tm.time()-tt)
```

```
Time take to retrieve every 2nd element from a list = 0.03579902648925781
Time take to retrieve every 2nd element from a tuple = 0.02684164047241211
```

2.2 Dictionary

A dictionary consists of key-value pairs, where the keys and values are python objects. While values can be any python object, keys need to be immutable python objects, like strings, integers, tuples, etc. Thus, a list can be a value, but not a key, as elements of a list can be changed. A dictionary is defined using the keyword `dict` along with curly braces, colons to separate keys and values, and commas to separate elements of a dictionary:

```
dict_example = {'USA': 'Joe Biden', 'India': 'Narendra Modi', 'China': 'Xi Jinping'}
```

Elements of a dictionary can be retrieved by using the corresponding key.

```
dict_example['India']
```

```
'Narendra Modi'
```

2.2.1 Adding and removing elements in a dictionary

New elements can be added to a dictionary by defining a key in square brackets and assigning it to a value:

```
dict_example['Japan'] = 'Fumio Kishida'
dict_example['Countries'] = 4
dict_example
```

```
{'USA': 'Joe Biden',
 'India': 'Narendra Modi',
 'China': 'Xi Jinping',
 'Japan': 'Fumio Kishida',
 'Countries': 4}
```

Elements can be removed from the dictionary using the `del` method or the `pop` method:

```
#Removing the element having key as 'Countries'
del dict_example['Countries']
```



```
dict_example
```

```
{'USA': 'Joe Biden',  
 'India': 'Narendra Modi',  
 'China': 'Xi Jinping',  
 'Japan': 'Fumio Kishida'}
```

```
#Removing the element having key as 'USA'  
dict_example.pop('USA')
```

```
'Joe Biden'
```

```
dict_example
```

```
{'India': 'Narendra Modi', 'China': 'Xi Jinping', 'Japan': 'Fumio Kishida'}
```

New elements can be added, and values of existing keys can be changed using the `update` method:

```
dict_example = {'USA': 'Joe Biden', 'India': 'Narendra Modi', 'China': 'Xi Jinping', 'Countries': 3}  
dict_example
```

```
{'USA': ['Joe Biden'],  
 'India': 'Narendra Modi',  
 'China': 'Xi Jinping',  
 'Countries': 3}
```

```
dict_example.update({'Countries': 4, 'Japan': 'Fumio Kishida'})
```

```
dict_example
```

```
{'USA': 'Joe Biden',  
 'India': 'Narendra Modi',  
 'China': 'Xi Jinping',  
 'Countries': 4,  
 'Japan': 'Fumio Kishida'}
```

2.3 Functions

If an algorithm or block of code is being used several times in a code, then it can be separately defined as a function. This makes the code more organized and readable. For example, let us define a function that prints prime numbers between **a** and **b**, and returns the number of prime numbers found.

```
#Function definition
def prime_numbers (a,b=100):
    num_prime_nos = 0

    #Iterating over all numbers between a and b
    for i in range(a,b):
        num_divisors=0

        #Checking if the ith number has any factors
        for j in range(2, i):
            if i%j == 0:
                num_divisors=1;break;

        #If there are no factors, then printing and counting the number as prime
        if num_divisors==0:
            print(i)
            num_prime_nos = num_prime_nos+1

    #Return count of the number of prime numbers
    return num_prime_nos
```

In the above function, the keyword **def** is used to define the function, **prime_numbers** is the name of the function, **a** and **b** are the arguments that the function uses to compute the output.

Let us use the defined function to print and count the prime numbers between 40 and 60.

```
#Printing prime numbers between 40 and 60
num_prime_nos_found = prime_numbers(40,60)
```

41
43
47
53
59

```
num_prime_nos_found
```

5

If the user calls the function without specifying the value of the argument **b**, then it will take the default value of 100, as mentioned in the function definition. However, for the argument **a**, the user will need to specify a value, as there is no value defined as a default value in the function definition.

2.3.1 Global and local variables with respect to a function

A variable defined within a function is local to that function, while a variable defined outside the function is global with respect to that function. In case a variable with the same name is defined both outside and inside a function, it will refer to its global or local value, depending on where it occurs.

The example below shows a variable with the name **var** referring to its local value when called within the function, and global value when called outside the function.

```
var = 5
def sample_function(var):
    print("Local value of 'var' within 'sample_function()' = ",var)

sample_function(4)
print("Global value of 'var' outside 'sample_function()' = ",var)
```

Local value of 'var' within 'sample_function()' = 4

Global value of 'var' outside 'sample_function()' = 5

References