P4M Notebook 3

November 1, 2018

1 Part 3: Lists

Python provides a powerful set of tools to create and manipulate lists of data. In this part, we take a deep dive into the Python list type. This includes mutability, list methods, and slicing.

We will use Python lists to understand permutations, understanding the sign of a permutation in terms of transpositions, cycle-type, and inversions.

Then we use Python lists to implement and optimize the Sieve of Eratosthenes, which will produce a list of all prime numbers up to a big number (like 10 million) in a snap. Along the way, we introduce some Python techniques for data analysis and visualization.

1.1 Table of Contents

- Section **??**
- Section **??**
- Section **??**
- Section **??**

1.2 Primality testing

Before diving into lists, we recall the **brute force** primality test that we created in the last lesson. To test whether a number n is prime, we can simply check for factors. This yields the following primality test.

```
In [11]: def is_prime(n):
             '''
             Checks whether the argument n is a prime number.
             Uses a brute force search for factors between 1 and n.
             '''
             for j in range(2,n):  # the range of numbers 2,3,...,n-1.
                 if n%j == 0:  # is n divisible by j?
                     print("{} is a factor of {}.".format(j,n))
                     return False
             return True
```

We can also implement this test with a **while loop** instead of a for loop. This doesn't make much of a difference, in Python 3.x. (In Python 2.x, this would save memory).

```
In [12]: def is_prime(n):
             '''
             Checks whether the argument n is a prime number.
             Uses a brute force search for factors between 1 and n.
             '''
             j = 2
             while j < n:  # j will proceed through the list of numbers 2,3,...,n-1.
                 if n%j == 0:  # is n divisible by j?
                     print("{} is a factor of {}.".format(j,n))
                     return False
                 j = j + 1  # There's a Python abbreviation for this:  j += 1.
             return True
```

```
In [13]: is_prime(10001)
```

```
73 is a factor of 10001.
```

```
Out[13]: False
```

```
In [14]: is_prime(101)
```

```
Out[14]: True
```

If n is a prime number, then the `is_prime(n)` function will iterate through all the numbers between 2 and $n-1$. But this is overkill! Indeed, if n is not prime, it will have a factor between 2 and the square root of n. This is because factors come in pairs: if $ab = n$, then one of the factors, a or b, must be less than or equal to the square root of n. So it suffices to search for factors up to (and including) the square root of n.

Even though we've made our own sqrt function, we load a fast one from the standard math package. You can use this for square roots, trig functions, logs, and more. Click the previous link for documentation. This package doesn't load automatically when you start Python, so you have to load it with a little Python code.

```
In [15]: from math import sqrt
```

This command **imports** the square root function (`sqrt`) from the **package** called `math`. Now you can find square roots.

```
In [16]: sqrt(1000)
```

```
Out[16]: 31.622776601683793
```

There are a few different ways to import functions from packages. The above syntax is a good starting point, but sometimes problems can arise if different packages have functions with the same name. Here are a few methods of importing the `sqrt` function and how they differ.

`from math import sqrt`: After this command, `sqrt` will refer to the function from the `math` package (overriding any previous definition).

`import math`: After this command, all the functions from the `math` package will be imported. But to call `sqrt`, you would type a command like `math.sqrt(1000)`. This is convenient if there are potential conflicts with other packages.

`from math import *`: After this command, all the functions from the `math` package will be imported. To call them, you can access them directly with a command like `sqrt(1000)`. This can easily cause conflicts with other packages, since packages can have hundreds of functions in them!

`import math as mth`: Some people like abbreviations. This imports all the functions from the `math` package. To call one, you type a command like `mth.sqrt(1000)`.

```
In [17]: import math

In [18]: math.sqrt(1000)

Out[18]: 31.622776601683793

In [19]: factorial(10)  # This will cause an error!


        ---------------------------------------------------------------------------

        NameError                                 Traceback (most recent call last)

        <ipython-input-19-e379444d4994> in <module>()
    ----> 1 factorial(10)  # This will cause an error!


        NameError: name 'factorial' is not defined


In [ ]: math.factorial(10)  # This is ok, since the math package comes with a function called
```

Now let's improve our `is_prime(n)` function by searching for factors only up to the square root of the number `n`. We consider two options.

```
In [20]: def is_prime_slow(n):
             '''
             Checks whether the argument n is a prime number.
             Uses a brute force search for factors between 1 and n.
             '''
             j = 2
             while j <= sqrt(n):  # j will proceed through the list of numbers 2,3,... up to s
                 if n%j == 0:  # is n divisible by j?
                     print("{} is a factor of {}.".format(j,n))
                     return False
                 j = j + 1  # There's a Python abbreviation for this:  j += 1.
             return True

In [21]: def is_prime_fast(n):
             '''
             Checks whether the argument n is a prime number.
             Uses a brute force search for factors between 1 and n.
             '''
```

```
        j = 2
        root_n = sqrt(n)
        while j <= root_n:  # j will proceed through the list of numbers 2,3,... up to sq
            if n%j == 0:  # is n divisible by j?
                print("{} is a factor of {}.".format(j,n))
                return False
            j = j + 1  # There's a Python abbreviation for this:  j += 1.
        return True
```

```
In [22]: is_prime_fast(1000003)

Out[22]: True
```

```
In [23]: is_prime_slow(1000003)

Out[23]: True
```

I've chosen function names with "fast" and "slow" in them. But what makes them faster or slower? Are they faster than the original? And how can we tell?

Python comes with a great set of tools for these questions. The simplest (for the user) are the time utilities. By placing the **magic** `%timeit` before a command, Python does something like the following:

1. Python makes a little container in your computer devoted to the computations, to avoid interference from other running programs if possible.
2. Python executes the command lots and lots of times.
3. Python averages the amount of time taken for each execution.

Give it a try below, to compare the speed of the functions `is_prime` (the original) with the new `is_prime_fast` and `is_prime_slow`. Note that the `%timeit` commands might take a little while.

```
In [24]: %timeit is_prime_fast(1000003)

120 ţs ś 339 ns per loop (mean ś std. dev. of 7 runs, 10000 loops each)
```

```
In [25]: %timeit is_prime_slow(1000003)

188 ţs ś 1.53 ţs per loop (mean ś std. dev. of 7 runs, 1000 loops each)
```

```
In [26]: %timeit is_prime(1000003)

114 ms ś 3.38 ms per loop (mean ś std. dev. of 7 runs, 10 loops each)
```

Time is measured in seconds, milliseconds (1 ms = 1/1000 second), microseconds (1 ţs = 1/1,000,000 second), and nanoseconds (1 ns = 1/1,000,000,000 second). So it might appear at first that `is_prime` is the fastest, or about the same speed. But check the units! The other two approaches are about a thousand times faster! How much faster were they on your computer?

```
In [ ]: is_prime_fast(10000000000037)  # Don't try this with `is_prime` unless you want to wai
```

Indeed, the `is_prime_fast(n)` function will go through a loop of length about `sqrt(n)` when `n` is prime. But `is_prime(n)` will go through a loop of length about `n`. Since `sqrt(n)` is much less than `n`, especially when `n` is large, the `is_prime_fast(n)` function is much faster.

Between `is_prime_fast` and `is_prime_slow`, the difference is that the `fast` version **precomputes** the square root `sqrt(n)` before going through the loop, where the `slow` version repeats the `sqrt(n)` every time the loop is repeated. Indeed, writing `while j <= sqrt(n):` suggests that Python might execute `sqrt(n)` every time to check. This *might* lead to Python computing the same square root a million times... unnecessarily!

A basic principle of programming is to **avoid repetition**. If you have the memory space, just compute once and store the result. It will probably be faster to pull the result out of memory than to compute it again.

Python does tend to be pretty smart, however. It's possible that Python **is precomputing** `sqrt(n)` even in the slow loop, just because it's clever enough to tell in advance that the same thing is being computed over and over again. This depends on your Python version and takes place behind the scenes. If you want to figure it out, there's a whole set of tools (for advanced programmers) like the disassembler to figure out what Python is doing.

If you feel like looking under the hood, the next few lines will display the `is_prime_fast` and `is_prime_slow` functions to bytecode. Can you see how the `sqrt(n)` computation is carried out differently?

```
In [27]: from dis import dis
```

```
In [28]: dis(is_prime_fast)
```

```
  6           0 LOAD_CONST               1 (2)
              2 STORE_FAST               1 (j)

  7           4 LOAD_GLOBAL              0 (sqrt)
              6 LOAD_FAST                0 (n)
              8 CALL_FUNCTION            1
             10 STORE_FAST               2 (root_n)

  8          12 SETUP_LOOP              52 (to 66)
        >>   14 LOAD_FAST                1 (j)
             16 LOAD_FAST                2 (root_n)
             18 COMPARE_OP               1 (<=)
             20 POP_JUMP_IF_FALSE       64

  9          22 LOAD_FAST                0 (n)
             24 LOAD_FAST                1 (j)
             26 BINARY_MODULO
             28 LOAD_CONST               2 (0)
             30 COMPARE_OP               2 (==)
             32 POP_JUMP_IF_FALSE       54

 10          34 LOAD_GLOBAL              1 (print)
             36 LOAD_CONST               3 ('{} is a factor of {}.')
             38 LOAD_ATTR                2 (format)
             40 LOAD_FAST                1 (j)
             42 LOAD_FAST                0 (n)
             44 CALL_FUNCTION            2
             46 CALL_FUNCTION            1
             48 POP_TOP

 11          50 LOAD_CONST               4 (False)
             52 RETURN_VALUE

 12     >>   54 LOAD_FAST                1 (j)
             56 LOAD_CONST               5 (1)
             58 BINARY_ADD
             60 STORE_FAST               1 (j)
             62 JUMP_ABSOLUTE           14
        >>   64 POP_BLOCK

 13     >>   66 LOAD_CONST               6 (True)
             68 RETURN_VALUE
```

```
In [29]: dis(is_prime_slow)
```

```
  6           0 LOAD_CONST               1 (2)
              2 STORE_FAST               1 (j)

  7           4 SETUP_LOOP              56 (to 62)
        >>    6 LOAD_FAST                1 (j)
              8 LOAD_GLOBAL              0 (sqrt)
             10 LOAD_FAST                0 (n)
             12 CALL_FUNCTION            1
             14 COMPARE_OP               1 (<=)
             16 POP_JUMP_IF_FALSE       60

  8          18 LOAD_FAST                0 (n)
             20 LOAD_FAST                1 (j)
             22 BINARY_MODULO
             24 LOAD_CONST               2 (0)
             26 COMPARE_OP               2 (==)
             28 POP_JUMP_IF_FALSE       50

  9          30 LOAD_GLOBAL              1 (print)
             32 LOAD_CONST               3 ('{} is a factor of {}.')
             34 LOAD_ATTR                2 (format)
             36 LOAD_FAST                1 (j)
             38 LOAD_FAST                0 (n)
             40 CALL_FUNCTION            2
             42 CALL_FUNCTION            1
             44 POP_TOP

 10          46 LOAD_CONST               4 (False)
             48 RETURN_VALUE

 11     >>   50 LOAD_FAST                1 (j)
             52 LOAD_CONST               5 (1)
             54 BINARY_ADD
             56 STORE_FAST               1 (j)
             58 JUMP_ABSOLUTE            6
        >>   60 POP_BLOCK

 12     >>   62 LOAD_CONST               6 (True)
             64 RETURN_VALUE
```

```
In [30]: is_prime_fast(10**14 + 37) # This might get a bit of delay.
```

```
1858741 is a factor of 100000000000037.
```

```
Out[30]: False
```

Now we have a function `is_prime_fast(n)` that is speedy for numbers `n` in the trillions! You'll probably start to hit a delay around 10^{15} or so, and the delays will become intolerable if you add too many more digits. In a future lesson, we will see a different primality test that will be essentially instant even for numbers around 10^{1000}!

1.2.1 Exercises

1. To check whether a number `n` is prime, you can first check whether `n` is even, and then check whether `n` has any odd factors. Change the `is_prime_fast` function by implementing this improvement. How much of a speedup did you get?

2. Use the `%timeit` tool to study the speed of `is_prime_fast` for various sizes of `n`. Using about 10 data points, relate the size of `n` to the time taken by the `is_prime_fast` function.

3. Write a function `is_square(n)` to test whether a given integer `n` is a perfect square (like 0, 1, 4, 9, 16, etc.). How fast can you make it run? Describe the different approaches you try and which are fastest.

```
In [31]: #1
         def is_prime_fast(n):
             # first check if n is even
             # if yes, return False
             if(n%2 == 0):
                 return False
             else:
                 j = 3
                 root_n = sqrt(n)
```

```
            while j <= root_n:  # j will proceed through the list of numbers 2,3,... up t
                if n%j == 0:  # is n divisible by j?
                    #print("{} is a factor of {}.".format(j,n))
                    return False
                j += 2  # There's a Python abbreviation for this:  j += 1.
        return True
```

```
In [32]: %timeit is_prime_fast(1000003)

60.7 ţs ś 352 ns per loop (mean ś std. dev. of 7 runs, 10000 loops each)
```

This is nearly twice as fast as the old is_prime_fast() function.

```
In [33]: %timeit is_prime_fast(10000007)

57.7 ţs ś 956 ns per loop (mean ś std. dev. of 7 runs, 10000 loops each)
```

```
In [34]: %timeit is_prime_fast(100000007)

645 ţs ś 11.7 ţs per loop (mean ś std. dev. of 7 runs, 1000 loops each)
```

```
In [35]: %timeit is_prime_fast(1000000007)

2.03 ms ś 22.2 ţs per loop (mean ś std. dev. of 7 runs, 100 loops each)
```

```
In [36]: %timeit is_prime_fast(10000000007)

1.94 ţs ś 28.7 ns per loop (mean ś std. dev. of 7 runs, 100000 loops each)
```

```
In [37]: %timeit is_prime_fast(100000000007)

26.7 ţs ś 230 ns per loop (mean ś std. dev. of 7 runs, 10000 loops each)
```

```
In [38]: %timeit is_prime_fast(1000000000007)

3.42 ms ś 780 ţs per loop (mean ś std. dev. of 7 runs, 100 loops each)
```

Prime numbers take much longer to test for obvious reasons. The size of the number doesn't really seem to have a huge impact on the time it takes, what has more impact is what the smallest factor of the large number is because we have to test to that number.

```
In [39]: #3
         import math
         def is_square_method_1(n):
             # check if the square root of n is an integer
             # if True then it's a perfect square
             if(math.sqrt(n).is_integer()):
                 return True
             else:
                 return False

In [40]: %timeit is_square_method_1(12723679885609870147705078125)

268 ns ś 12.5 ns per loop (mean ś std. dev. of 7 runs, 1000000 loops each)


In [41]: def is_square_method_2(n):
             square = math.sqrt(n)
             # check if the square of n is an odd or even number (indirectly checking if it's
             if((square%2 == 0) or (square%2 == 1)):
                 return True
             else:
                 return False


In [42]: %timeit is_square_method_2(12723679885609870147705078125)

498 ns ś 8.32 ns per loop (mean ś std. dev. of 7 runs, 1000000 loops each)


In [43]: def is_square_method_3(n):
             # this is a really bad way of doing it but I really want to time it
             # have to make sure it's an 'int' to use the range() function
             square = int(math.sqrt(n))
             for x in range(square):
                 if(x**2 == n):
                     return True
             return False

In [44]: %timeit is_square_method_3(12723679885609870147705078125)
         # this was taking too long so I had to interrupt the kernal


        ---------------------------------------------------------------------------

        KeyboardInterrupt                         Traceback (most recent call last)

        <ipython-input-44-f97e7d0ab512> in <module>()
    ----> 1 get_ipython().run_line_magic('timeit', 'is_square_method_3(12723679885609870147705
          2 # this was taking too long so I had to interrupt the kernal
        ~/anaconda3/lib/python3.6/site-packages/IPython/core/interactiveshell.py in run_line_ma
       2129                 kwargs['local_ns'] = sys._getframe(stack_depth).f_locals
       2130             with self.builtin_trap:
    -> 2131                 result = fn(*args,**kwargs)
       2132             return result
       2133 


        <decorator-gen-61> in timeit(self, line, cell, local_ns)


        ~/anaconda3/lib/python3.6/site-packages/IPython/core/magic.py in <lambda>(f, *a, **k)
        185     # but it's overkill for just that one bit of state.
        186     def magic_deco(arg):
    --> 187         call = lambda f, *a, **k: f(*a, **k)
        188 
        189         if callable(arg):


        ~/anaconda3/lib/python3.6/site-packages/IPython/core/magics/execution.py in timeit(sel
       1096             for index in range(0, 10):
       1097                 number = 10 ** index
    -> 1098                 time_number = timer.timeit(number)
       1099                 if time_number >= 0.2:
       1100                     break


        ~/anaconda3/lib/python3.6/site-packages/IPython/core/magics/execution.py in timeit(sel
        158         gc.disable()
        159         try:
    --> 160             timing = self.inner(it, self.timer)
        161         finally:
        162             if gcold:


        <magic-timeit> in inner(_it, _timer)


        <ipython-input-43-1233654b3109> in is_square_method_3(n)
          4     square = int(math.sqrt(n))
          5     for x in range(square):
    ----> 6         if(x**2 == n):
          7             return True
          8     return False
```

```
        KeyboardInterrupt:
```

1.3 List manipulation

We have already (briefly) encountered the `list` type in Python. Recall that the `range` command produces a range, which can be used to produce a list. For example, `list(range(10))` produces the list `[0,1,2,3,4,5,6,7,8,9]`. You can also create your own list by a writing out its terms, e.g. `L = [4,7,10]`.

Here we work with lists, and a very Pythonic approach to list manipulation. With practice, this can be a powerful tool to write fast algorithms, exploiting the hard-wired capability of your computer to shift and slice large chunks of data. Our eventual application will be to implement the Sieve of Eratosthenes, producing a long list of prime numbers (without using any `is_prime` test along the way).

We begin by creating a list to play with. We mix numbers and strings... just for fun.

```
In [85]: L = [0,'one',2,'three',4,'five',6,'seven',8,'nine',10]
```

1.3.1 List terms and indices

Notice that the entries in a list can be of any type. The above list `L` has some integer entries and some string entries. Lists are **ordered** in Python, **starting at zero**. One can access the n^{th} entry in a list with a command like `L[n]`.

```
In [86]: L[3]

Out[86]: 'three'

In [87]: print(L[3])  # Note that Python has slightly different approaches to the print-functi

three


In [88]: print(L[4])  # We will use the print function, because it makes our printing intentio

4


In [89]: print(L[0])

0
```

The location of an entry is called its **index**. So *at* the index 3, the list `L` stores the entry `three`. Note that the same entry can occur in many places in a list. E.g. `[7,7,7]` is a list with 7 at the zeroth, first, and second index.

```
In [90]: print(L[-1])
         print(L[-2])
```

```
10
nine
```

The last bit of code demonstrates a cool Python trick. The "-1st" entry in a list refers to the last entry. The "-2nd entry" refers to the second-to-last entry, and so on. It gives a convenient way to access both sides of the list, even if you don't know how long it is.

Of course, you can use Python to find out how long a list is.

```
In [91]: len(L)
```

```
Out[91]: 11
```

You can also use Python to find the sum of a list of numbers.

```
In [92]: sum([1,2,3,4,5])
```

```
Out[92]: 15
```

```
In [93]: sum(range(100))  # Be careful.  This is the sum of which numbers?  # The sum function
```

```
Out[93]: 4950
```

1.3.2 List slicing

Slicing lists allows us to create new lists (or ranges) from old lists (or ranges), by chopping off one end or the other, or even slicing out entries at a fixed interval. The simplest syntax has the form `L[a:b]` where `a` denotes the index of the starting entry and index of the final entry is one less than `b`. It is best to try a few examples to get a feel for it.

Slicing a list with a command like `L[a:b]` doesn't actually *change* the original list `L`. It just extracts some terms from the list and outputs those terms. Soon enough, we will change the list `L` using a list assignment.

```
In [94]: L[0:5]
```

```
Out[94]: [0, 'one', 2, 'three', 4]
```

```
In [95]: L[5:11]  # Notice that L[0:5] and L[5:11] together recover the whole list.
```

```
Out[95]: ['five', 6, 'seven', 8, 'nine', 10]
```

```
In [96]: L[3:7]
```

```
Out[96]: ['three', 4, 'five', 6]
```

This continues the strange (for beginners) Python convention of starting at the first number and ending just before the last number. Compare to `range(3,7)`, for example.

The command `L[0:5]` can be replaced by `L[:5]` to abbreviate. The empty opening index tells Python to start at the beginning. Similarly, the command `L[5:11]` can be replaced by `L[5:]`. The empty closing index tells Python to end the slice and the end. This is helpful if one doesn't know where the list ends.

```
In [97]: L[:5]

Out[97]: [0, 'one', 2, 'three', 4]

In [98]: L[3:]

Out[98]: ['three', 4, 'five', 6, 'seven', 8, 'nine', 10]
```

Just like the `range` command, list slicing can take an optional third argument to give a step size. To understand this, try the command below.

```
In [99]: L[2:10]

Out[99]: [2, 'three', 4, 'five', 6, 'seven', 8, 'nine']

In [100]: L[2:10:3]

Out[100]: [2, 'five', 8]
```

If, in this three-argument syntax, the first or second argument is absent, then the slice starts at the beginning of the list or ends at the end of the list accordingly.

```
In [101]: L  # Just a reminder.  We haven't modified the original list!

Out[101]: [0, 'one', 2, 'three', 4, 'five', 6, 'seven', 8, 'nine', 10]

In [102]: L[:9:3]   # Start at zero, go up to (but not including) 9, by steps of 3.

Out[102]: [0, 'three', 6]

In [103]: L[2: :3] # Start at two, go up through the end of the list, by steps of 3.

Out[103]: [2, 'five', 8]

In [104]: L[::3]  # Start at zero, go up through the end of the list, by steps of 3.

Out[104]: [0, 'three', 6, 'nine']
```

1.3.3 Changing list slices

Not only can we extract and study terms or slices of a list, we can change them by assignment. The simplest case would be changing a single term of a list.

```
In [105]: print(L) # Start with the list L.

[0, 'one', 2, 'three', 4, 'five', 6, 'seven', 8, 'nine', 10]


In [106]: L[5] = 'Bacon!'

In [107]: print(L)  # What do you think L is now?
```

```
[0, 'one', 2, 'three', 4, 'Bacon!', 6, 'seven', 8, 'nine', 10]
```

```
In [108]: print(L[2::3]) # What do you think this will do?
```

```
[2, 'Bacon!', 8]
```

We can change an entire slice of a list with a single assignment. Let's change the first two terms of `L` in one line.

```
In [109]: L[:2] = ['Pancakes', 'Ham']  # What was L[:2] before?
```

```
In [110]: print(L) # Oh... what have we done!
```

```
['Pancakes', 'Ham', 2, 'three', 4, 'Bacon!', 6, 'seven', 8, 'nine', 10]
```

```
In [111]: L[0]
```

```
Out[111]: 'Pancakes'
```

```
In [112]: L[1]
```

```
Out[112]: 'Ham'
```

```
In [113]: L[2]
```

```
Out[113]: 2
```

We can change a slice of a list with a single assignment, even when that slice does not consist of consecutive terms. Try to predict what the following commands will do.

```
In [114]: print(L)  # Let's see what the list looks like before.
```

```
['Pancakes', 'Ham', 2, 'three', 4, 'Bacon!', 6, 'seven', 8, 'nine', 10]
```

```
In [115]: L[::2] = ['A','B','C','D','E','F']  # What was L[::2] before this assignment?
```

```
In [116]: print(L)  # What do you predict?
```

```
['A', 'Ham', 'B', 'three', 'C', 'Bacon!', 'D', 'seven', 'E', 'nine', 'F']
```

1.4 List methods

A method is a function that is attached to an object. We have already used one method: the `format` method that is attached to all strings. You might have seen the `replace` method for strings too. Note that single-quotes `'Hello'` or double-quotes `"Hello"` can be used for strings.

```
In [117]: "Hello {}!".format('programming student')

Out[117]: 'Hello programming student!'

In [118]: "Programming is fun!".replace('fun','lit')

Out[118]: 'Programming is lit!'
```

List methods are functions attached to lists. Some useful methods include `append` and `sort`. A fuller listing can be found at the official documentation.

```
In [119]: L = [1,2,3]
          L.append(4)
          print(L)

[1, 2, 3, 4]
```

The `append` method can be used to add new items to the end of a list. But be careful if you want to add multiple items!

```
In [120]: L.append([5,6])
          print(L)

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

Behind the scenes, methods are functions which have a special input parameter called `self`. So when you use a command like `L.append(4)`, you are effectively running `append(L, 4)`. The `self` parameter is the object the method is attached to.

Like all functions, methods have outputs too. But what can be confusing is that methods can *modify* `self` and can sometimes *return* `None`.

```
In [121]: print([1,2,3].append(4))
          print("123".replace("3","4"))

None
124
```

This is very confusing at first! The list `append` method *does* change `self` by appending something to `self`. But as a function, it returns `None`.

On the other hand, the string `replace` method *does not* change `self` and instead *returns* the modified string.

This will make more sense after we study *mutable* and *immutable* types. Lists are mutable (and thus are often changed by their methods). Strings are immutable, and so changes are effected by producing new strings. Another example of a string method is `sort()`. The only parameter of `sort` is `self`, and so nothing needs to go between the paraentheses.

```
In [122]: L = [4,2,1]  # Make a list.
          L.sort()  # Sort the list.  This *changes* L and returns None.
          print(L)  # Let's see what L is now.
```

```
[1, 2, 4]
```

```
In [123]: L = ['Ukelele', 'Apple', 'Dog', 'Cat' ]
          L.sort()
          print(L)
```

```
['Apple', 'Cat', 'Dog', 'Ukelele']
```

Sorting numbers is possible, because the Python operator < is defined for numbers. Sorting strings is possible, because the Python operator < is interpreted alphabetically among strings. If you mix types, Python might not know how to behave... you'll get a TypeError.

```
In [124]: L = [1,'Apple', 3.14]
          L.sort()
          print(L)
```

```
        ---------------------------------------------------------------------------

        TypeError                                 Traceback (most recent call last)

        <ipython-input-124-e3d2589f72dc> in <module>()
          1 L = [1,'Apple', 3.14]
    ----> 2 L.sort()
          3 print(L)


        TypeError: '<' not supported between instances of 'str' and 'int'
```

1.4.1 Exercises

1. Create a list L with L = [1,2,3,...,100] (all the numbers from 1 to 100). What is L[50]?
2. Take the same list L, and extract a slice of the form [5,10,15,...,95] with a command of the form L[a:b:c].
3. Take the same list L, and change all the even numbers to zeros, so that L looks like [1,0,3,0,5,0,...,99,0]. Hint: You might wish to use the list [0]*50.
4. Try the command `L[-1::-1]` on a list. What does it do? Can you guess before executing it? Can you understand why? In fact, strings are indexed like lists. Try setting `L = 'Hello'` and the previous command.

5. Create the list [1,100,3,98,5,96,...,99,0], where the odd terms are in order and the even terms are in reverse order. There are multiple methods!
6. Use the append method with a loop to create a list of perfect squares, [0,1,4,9,16,25,...,10000].

```
In [125]: B = list(range(9,100,6))
          print(B)
          print(len(B))

[9, 15, 21, 27, 33, 39, 45, 51, 57, 63, 69, 75, 81, 87, 93, 99]
16
```

```
In [126]: #1
          L = list(range(1,101))

In [127]: #2
          print(L[4:95:5])

[5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95]
```

```
In [128]: #3
          for x in range(1, len(L), 2):
              L[x] = 0
          print(L)

[1, 0, 3, 0, 5, 0, 7, 0, 9, 0, 11, 0, 13, 0, 15, 0, 17, 0, 19, 0, 21, 0, 23, 0, 25, 0, 27, 0,
```

```
In [129]: #4
          L[-1::-1]
          # this starts at the last number and decrements by 1

Out[129]: [0,
           99,
           0,
           97,
           0,
           95,
           0,
           93,
           0,
           91,
           0,
           89,
           0,
           87,
           0,
           85,
0,
83,
0,
81,
0,
79,
0,
77,
0,
75,
0,
73,
0,
71,
0,
69,
0,
67,
0,
65,
0,
63,
0,
61,
0,
59,
0,
57,
0,
55,
0,
53,
0,
51,
0,
49,
0,
47,
0,
45,
0,
43,
0,
41,
0,
39,
0,
37,
 0,
 35,
 0,
 33,
 0,
 31,
 0,
 29,
 0,
 27,
 0,
 25,
 0,
 23,
 0,
 21,
 0,
 19,
 0,
 17,
 0,
 15,
 0,
 13,
 0,
 11,
 0,
 9,
 0,
 7,
 0,
 5,
 0,
 3,
 0,
 1]
```

```
In [130]: L = 'Hello'
          L[-1::-1]

Out[130]: 'olleH'

In [131]: # Create the list [1,100,3,98,5,96,...,99,0], where the odd terms are in order and th
          L = list(range(1,101))
          for x in range(1, len(L), 2):
              L[x] = 101-x
          print(L)

[1, 100, 3, 98, 5, 96, 7, 94, 9, 92, 11, 90, 13, 88, 15, 86, 17, 84, 19, 82, 21, 80, 23, 78, 25
```

```
In [132]: # Use the append method with a loop to create a list of perfect squares, [0,1,4,9,16
          perfect_squares = []
          for x in range(10001):
              if(is_square_method_1(x)):
                  perfect_squares.append(x)
          print(len(perfect_squares))
          print(perfect_squares)
```

```
101
[0, 1, 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400, 44
```

1.5 Sieve of Eratosthenes

The **Sieve of Eratosthenes** (hereafter called "the sieve") is a very fast way of producing long lists of primes, without doing repeated primality checking. The basic idea is to start with all of the natural numbers, and successively filter out, or **sieve**, the multiples of 2, then the multiples of 3, then the multiples of 5, etc., until only primes are left. You can read more about the sieve, and experimental number theory, at The Conversation

Using list slicing, we can carry out this sieving process efficiently. And with a few more tricks we encounter here, we can carry out the Sieve **very** efficiently.

1.5.1 The basic sieve

The first approach we introduce is a bit naive, but is a good starting place. We will begin with a list of numbers up to 100, and sieve out the appropriate multiples of 2,3,5,7.

```
In [133]: primes = list(range(100)) # Let's start with the numbers 0...99.
```

Now, to "filter", i.e., to say that a number is *not* prime, let's just change the number to the value `None`.

```
In [134]: primes[0] = None # Zero is not prime.
          primes[1] = None # One is not prime.
          print(primes) # What have we done?
```

```
[None, None, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 2
```

Now let's filter out the multiples of 2, starting at 4. This is the slice `primes[4::2]`

```
In [135]: primes[4::2] = [None] * len(primes[4::2])  # The right side is a list of Nones, of t
          print(primes) # What have we done?
```

```
[None, None, 2, 3, None, 5, None, 7, None, 9, None, 11, None, 13, None, 15, None, 17, None, 19
```

Now we filter out the multiples of 3, starting at 9.

```
In [136]: primes[9::3] = [None] * len(primes[9::3])  # The right side is a list of Nones, of t
          print(primes) # What have we done?
```

```
[None, None, 2, 3, None, 5, None, 7, None, None, None, 11, None, 13, None, None, None, 17, None
```

Next the multiples of 5, starting at 25 (the first multiple of 5 greater than 5 that's left!)

```
In [137]: primes[25::5] = [None] * len(primes[25::5])  # The right side is a list of Nones, of
          print(primes) # What have we done?
```

```
[None, None, 2, 3, None, 5, None, 7, None, None, None, 11, None, 13, None, None, None, 17, None
```

Finally, the multiples of 7, starting at 49 (the first multiple of 7 greater than 7 that's left!)

```
In [138]: primes[49::7] = [None] * len(primes[49::7])  # The right side is a list of Nones, of
          print(primes) # What have we done?
```

```
[None, None, 2, 3, None, 5, None, 7, None, None, None, 11, None, 13, None, None, None, 17, None
```

What's left? A lot of `None`s and the prime numbers up to 100. We have successfully sieved out all the nonprime numbers in the list, using just four sieving steps (and setting 0 and 1 to `None` manually).

But there's a lot of room for improvement, from beginning to end!

1. The format of the end result is not so nice.
2. We had to sieve each step manually. It would be much better to have a function `prime_list(n)` which would output a list of primes up to `n` without so much supervision.
3. The memory usage will be large, if we need to store all the numbers up to a large `n` at the beginning.

We solve these problems in the following way.

1. We will use a list of **booleans** rather than a list of numbers. The ending list will have a `True` value at prime indices and a `False` value at composite indices. This reduces the memory usage and increases the speed.

2. A `which` function (explained soon) will make the desired list of primes after everything else is done.
3. We will proceed through the sieving steps algorithmically rather than entering each step manually.

Here is a somewhat efficient implementation of the Sieve in Python.

```
In [139]: def isprime_list(n):
              '''
              Return a list of length n+1
              with Trues at prime indices and Falses at composite indices.
              '''
              flags = [True] * (n+1)  # A list [True, True, True,...] to start.
              flags[0] = False  # Zero is not prime.  So its flag is set to False.
              flags[1] = False  # One is not prime.  So its flag is set to False.
              p = 2  # The first prime is 2.  And we start sieving by multiples of 2.

              while p <= sqrt(n):  # We only need to sieve by p is p <= sqrt(n).
                  if flags[p]:  # We sieve the multiples of p if flags[p]=True.
                      flags[p*p::p] = [False] * len(flags[p*p::p]) # Sieves out multiples of p
                  p = p + 1 # Try the next value of p.

              return flags
```

```
In [140]: print(isprime_list(100))
```

```
[False, False, True, True, False, True, False, True, False, False, False, True, False, True, Fa
```

If you look carefully at the list of booleans, you will notice a `True` value at the 2nd index, the 3rd index, the 5th index, the 7th index, etc.. The indices where the values are `True` are precisely the **prime** indices. Since booleans take the smallest amount of memory of any data type (one **bit** of memory per boolean), your computer can carry out the `isprime_list(n)` function even when `n` is very large.

To be more precise, there are 8 bits in a **byte**. There are 1024 bytes (about 1000) in a kilobyte. There are 1024 kilobytes in a megabyte. There are 1024 megabytes in a gigabyte. Therefore, a gigabyte of memory is enough to store about 8 billion bits. That's enough to store the result of `isprime_list(n)` when `n` is about 8 billion. Not bad! And your computer probably has 4 or 8 or 12 or 16 gigabytes of memory to use.

To transform the list of booleans into a list of prime numbers, we create a function called `where`. This function uses another Python technique called **list comprehension**. We discuss this technique later in this lesson, so just use the `where` function as a tool for now, or read about list comprehension if you're curious.

```
In [149]: def where(L):
              '''
              Take a list of booleans as input and
              outputs the list of indices where True occurs.
              '''
              return [n for n in range(len(L)) if L[n]]
```

Combined with the `isprime_list` function, we can produce long lists of primes.

```
In [150]: print(where(isprime_list(100)))
```

```
[2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 9
```

Let's push it a bit further. How many primes are there between 1 and 1 million? We can figure this out in three steps:

1. Create the isprime_list.

2. Use where to get the list of primes.
3. Find the length of the list of primes.

But it's better to do it in two steps.

1. Create the isprime_list.
2. Sum the list! (Note that `True` is 1, for the purpose of summation!)

```
In [151]: sum(isprime_list(1000000))  # The number of primes up to a million!

Out[151]: 78498

In [152]: %timeit isprime_list(10**6)  # 1000 ms = 1 second.

45.2 ms ś 7.37 ms per loop (mean ś std. dev. of 7 runs, 10 loops each)


In [153]: %timeit sum(isprime_list(10**6))

67.4 ms ś 4.19 ms per loop (mean ś std. dev. of 7 runs, 10 loops each)
```

This isn't too bad! It takes a fraction of a second to identify the primes up to a million, and a smaller fraction of a second to count them! But we can do a little better.

The first improvement is to take care of the even numbers first. If we count carefully, then the sequence 4,6,8,...,n (ending at n-1 if n is odd) has the floor of (n-2)/2 terms. Thus the line `flags[4::2] = [False] * ((n-2)//2)` will set all the flags to False in the sequence 4,6,8,10,... From there, we can begin sieving by *odd* primes starting with 3.

The next improvement is that, since we've already sieved out all the even numbers (except 2), we don't have to sieve out again by *even multiples*. So when sieving by multiples of 3, we don't have to sieve out 9,12,15,18,21,etc.. We can just sieve out 9,15,21,etc.. When p is an odd prime, this can be taken care of with the code `flags[p*p::2*p] = [False] * len(flags[p*p::2*p])`.

```
In [263]: def isprime_list(n):
              '''
              Return a list of length n+1
              with Trues at prime indices and Falses at composite indices.
              '''
              flags = [True] * (n+1)  # A list [True, True, True,...] to start.
              flags[0] = False  # Zero is not prime.  So its flag is set to False.
              flags[1] = False  # One is not prime.  So its flag is set to False.
              flags[4::2] = [False] * ((n-2)//2)
              p = 3
              while p <= sqrt(n):  # We only need to sieve by p is p <= sqrt(n).
                  if flags[p]:  # We sieve the multiples of p if flags[p]=True.
                      if (len(flags[p*p::2*p]) != 0):
                          flags[p*p::2*p] = [False] * len(flags[p*p::2*p]) # Sieves out multip
                  p = p + 2 # Try the next value of p.  Note that we can proceed only through

              return flags
```

```
In [155]: %timeit sum(isprime_list(10**6))  # How much did this speed it up?

63.1 ms ś 640 ţs per loop (mean ś std. dev. of 7 runs, 10 loops each)
```

Another modest improvement is the following. In the code above, the program *counts* the terms in sequences like 9,15,21,27,..., in order to set them to `False`. This is accomplished with the length command `len(flags[p*p::2*p])`. But that length computation is a bit too intensive. A bit of algebraic work shows that the length is given formulaically in terms of `p` and `n` by the formula:

$$len = \lfloor \frac{n - p^2 - 1}{2p} \rfloor + 1$$

(Here $\lfloor x \rfloor$ denotes the floor function, i.e., the result of rounding down.) Putting this into the code yields the following.

```
In [156]: B = list(range(9:100:6))


          File "<ipython-input-156-9d2574329c47>", line 1
        B = list(range(9:100:6))
                        ^
    SyntaxError: invalid syntax



In [157]: def isprime_list(n):
              '''
              Return a list of length n+1
              with Trues at prime indices and Falses at composite indices.
              '''
              flags = [True] * (n+1)  # A list [True, True, True,...] to start.
              flags[0] = False  # Zero is not prime.  So its flag is set to False.
              flags[1] = False  # One is not prime.  So its flag is set to False.
              flags[4::2] = [False] * ((n-2)//2)
              p = 3
              while p <= sqrt(n):  # We only need to sieve by p is p <= sqrt(n).
                  if flags[p]:  # We sieve the multiples of p if flags[p]=True.
                      flags[p*p::2*p] = [False] * ((n-p*p-1)//(2*p)+1) # Sieves out multiples
                  p = p + 2 # Try the next value of p.

              return flags

In [158]: %timeit sum(isprime_list(10**6))  # How much did this speed it up?

48.2 ms ś 195 ţs per loop (mean ś std. dev. of 7 runs, 10 loops each)
```

That should be pretty fast! It should be under 100 ms (one tenth of one second!) to determine the primes up to a million, and on a newer computer it should be under 50ms. We have gotten pretty close to the fastest algorithms that you can find in Python, without using external packages (like SAGE or sympy). See the related discussion on StackOverflow... the code in this lesson was influenced by the code presented there.

1.5.2 Exercises

1. Prove that the length of `range(p*p, n, 2*p)` equals $\lfloor \frac{n-p^2-1}{2p} \rfloor + 1$.

2. A natural number n is called squarefree if it has no perfect square divides n except for 1. Write a function `squarefree_list(n)` which outputs a list of booleans: `True` if the index is squarefree and `False` if the index is not squarefree. For example, if you execute `squarefree_list(12)`, the output should be `[False, True, True, True, False, True, True, True, False, False, True, True, False]`. Note that the `False` entries are located the indices 0, 4, 8, 9, 12. These natural numbers have perfect square divisors besides 1.

3. Your DNA contains about 3 billion base pairs. Each "base pair" can be thought of as a letter, A, T, G, or C. How many bits would be required to store a single base pair? In other words, how might you convert a sequence of booleans into a letter A,T,G, or C? Given this, how many megabytes or gigabytes are required to store your DNA? How many people's DNA would fit on a thumb-drive?

```
In [159]: #2
          def squarefree_list(n):
              # create list of length n+1 with all entries set to True
              flags = [True] * (n+1)
              flags[0] = False
              # search from 2 to n+1 for perfect squares
              # once one is found, iterate through all multiples of it and set to False
              for s in range(2,n+1):
                  if(is_square_method_1(s)):
                      for x in range(s,n+1,s):
                          #if( x%s == 0):
                          flags[x] = False
              print(flags,'\n')
              return(flags)
```

```
In [160]: print(where(squarefree_list(12)))
```

```
[False, True, True, True, False, True, True, True, False, False, True, True, False]

[1, 2, 3, 5, 6, 7, 10, 11]
```

2 3 Since there are 4 possibilities for each we can encode them all using 2 bits since 2^2=4.

Since there are 3 billion base pairs and 2 values in a pair, we'll need 3 billion pairs * 2 bits to store them all. That turns out to be 6 billion bits, divide that by 8 to get 750,000,000 which is 0.75 gigabytes to store one person worth of DNA. My 16gb flashdrive could store 20 people on it.

2.1 Data analysis

Now that we can produce a list of prime numbers quickly, we can do some data analysis: some experimental number theory to look for trends or patterns in the sequence of prime numbers. Since Euclid (about 300 BCE), we have known that there are infinitely many prime numbers. But how are they distributed? What proportion of numbers are prime, and how does this proportion change over different ranges? As theoretical questions, these belong the the field of analytic number theory. But it is hard to know what to prove without doing a bit of experimentation. And so, at least since Gauss (read Tschinkel's article about Gauss's tables) started examining his extensive tables of prime numbers, mathematicians have been carrying out experimental number theory.

2.1.1 Analyzing the list of primes

Let's begin by creating our data set: the prime numbers up to 1 million.

```
In [161]: primes = where(isprime_list(1000000))

In [162]: len(primes) # Our population size.  A statistician might call it N.

Out[162]: 78498

In [163]: primes[-1]  # The last prime in our list, just before one million.

Out[163]: 999983

In [164]: type(primes) # What type is this data?

Out[164]: list

In [165]: print(primes[:100]) # The first hundred prime numbers.

[2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 9
```

To carry out serious analysis, we will use the method of **list comprehension** to place our population into "bins" for statistical analysis. Our first type of list comprehension has the form `[x for x in LIST if CONDITION]`. This produces the list of all elements of LIST satisfying CONDITION. It is similar to list slicing, except we pull out terms from the list according to whether a condition is true or false.

For example, let's divide the (odd) primes into two classes. Red primes will be those of the form 4n+1. Blue primes will be those of the form 4n+3. In other words, a prime `p` is red if `p%4 == 1` and blue if `p%4 == 3`. And the prime 2 is neither red nor blue.

```
In [166]: redprimes = [p for p in primes if p%4 == 1] # Note the [x for x in LIST if CONDITION]
          blueprimes = [p for p in primes if p%4 == 3]

          print('Red primes:',redprimes[:20]) # The first 20 red primes.
          print('Blue primes:',blueprimes[:20]) # The first 20 blue primes.

Red primes: [5, 13, 17, 29, 37, 41, 53, 61, 73, 89, 97, 101, 109, 113, 137, 149, 157, 173, 181
Blue primes: [3, 7, 11, 19, 23, 31, 43, 47, 59, 67, 71, 79, 83, 103, 107, 127, 131, 139, 151,
In [167]: print("There are {} red primes and {} blue primes, up to 1 million.".format(len(redpr

There are 39175 red primes and 39322 blue primes, up to 1 million.
```

This is pretty close! It seems like prime numbers are about evenly distributed between red and blue. Their remainder after division by 4 is about as likely to be 1 as it is to be 3. In fact, it is proven that *asymptotically* the ratio between the number of red primes and the number of blue primes approaches 1. However, Chebyshev noticed a persistent slight bias towards blue primes along the way.

Some of the deepest conjectures in mathematics relate to the prime counting function $\pi(x)$. Here $\pi(x)$ is the **number of primes** between 1 and x (inclusive). So $\pi(2) = 1$ and $\pi(3) = 2$ and $\pi(4) = 2$ and $\pi(5) = 3$. One can compute a value of $\pi(x)$ pretty easily using a list comprehension.

```
In [168]: def primes_upto(x):
              return len([p for p in primes if p <= x]) # List comprehension recovers the prim

In [169]: primes_upto(1000)  # There are 168 primes between 1 and 1000.

Out[169]: 168
```

Now we graph the prime counting function. To do this, we use a list comprehension, and the visualization library called matplotlib. For graphing a function, the basic idea is to create a list of x-values, a list of corresponding y-values (so the lists have to be the same length!), and then we feed the two lists into matplotlib to make the graph.

We begin by loading the necessary packages.

```
In [170]: import matplotlib  #  A powerful graphics package.
          import numpy  #  A math package
          import matplotlib.pyplot as plt  # A plotting subpackage in matplotlib.
```

Now let's graph the function $y = x^2$ over the domain $-2 \leq x \leq 2$ for practice. As a first step, we use numpy's `linspace` function to create an evenly spaced set of 11 x-values between -2 and 2.

```
In [171]: x_values = numpy.linspace(-2,2,11)  # The argument 11 is the *number* of terms, not
          print(x_values)
          type(x_values)

[-2.  -1.6 -1.2 -0.8 -0.4  0.   0.4  0.8  1.2  1.6  2. ]


Out[171]: numpy.ndarray
```

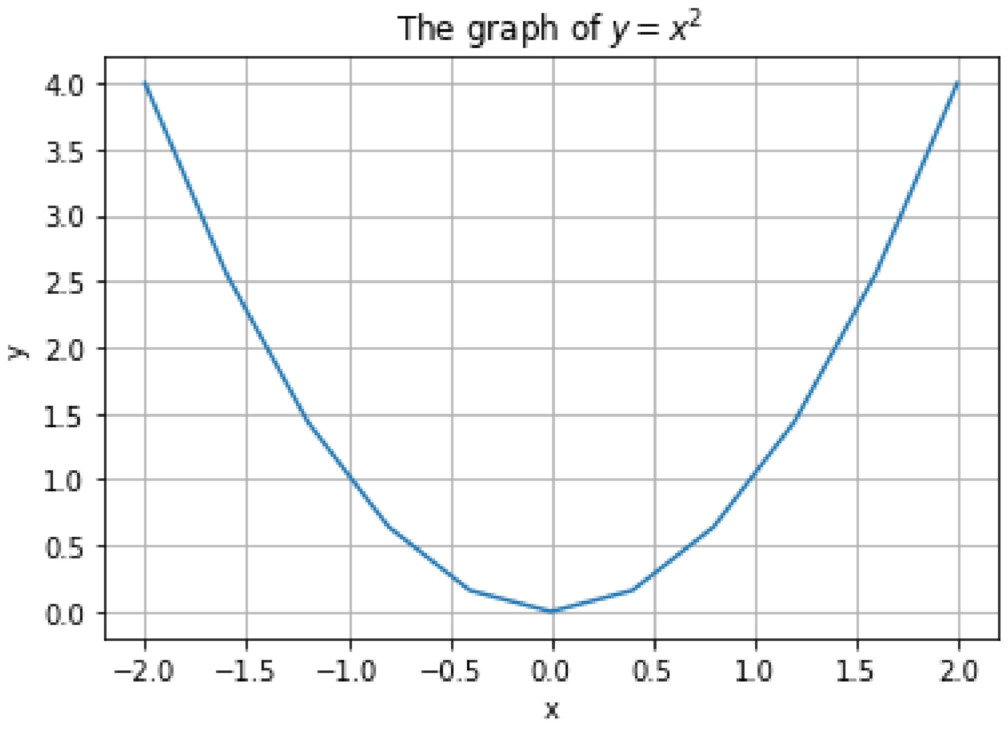
You might notice that the format looks a bit different from a list. Indeed, if you check `type(x_values)`, it's not a list but something else called a numpy array. Numpy is a package that excels with computations on large arrays of data. On the surface, it's not so different from a list. The `numpy.linspace` command is a convenient way of producing an evenly spaced list of inputs.

The big difference is that operations on numpy arrays are interpreted differently than operations on ordinary Python lists. Try the two commands for comparison.

```
In [172]: [1,2,3] + [1,2,3]

Out[172]: [1, 2, 3, 1, 2, 3]

In [173]: x_values + x_values

Out[173]: array([-4. , -3.2, -2.4, -1.6, -0.8,  0. ,  0.8,  1.6,  2.4,  3.2,  4. ])

In [174]: y_values = x_values * x_values  # How is multiplication interpreted on numpy arrays?
          print(y_values)

[4.   2.56 1.44 0.64 0.16 0.   0.16 0.64 1.44 2.56 4.  ]
```

Now we use matplotlib to create a simple line graph.

```
In [175]: %matplotlib inline
          plt.plot(x_values, y_values)
          plt.title('The graph of $y = x^2$')  # The dollar signs surround the formula, in LaT
          plt.ylabel('y')
          plt.xlabel('x')
          plt.grid(True)
          plt.show()
```



Let’s analyze the graphing code a bit more. See the official pyplot tutorial for more details.

```
%matplotlib inline
plt.plot(x_values, y_values)
plt.title('The graph of $y = x^2$')  # The dollar signs surround the formula, in LaTeX format.
plt.ylabel('y')
plt.xlabel('x')
plt.grid(True)
plt.show()
```

The first line contains the **magic** `%matplotlib inline`. We have seen a magic word before, in `%timeit`. Magic words can call another program to assist. So here, the magic `%matplotlib inline` calls matplotlib for help, and places the resulting figure within the notebook.

The next line `plt.plot(x_values, y_values)` creates a `plot object` based on the data of the x-values and y-values. It is an abstract sort of object, behind the scenes, in a format that matplotlib understands. The following lines set the title of the plot, the axis labels, and turns a grid on. The last line `plt.show` renders the plot as an image in your notebook. There's an infinite variety of graphs that matplotlib can produce -- see the gallery for more! Other graphics packages include bokeh and seaborn, which extends matplotlib.

2.1.2 Analysis of the prime counting function

Now, to analyze the prime counting function, let's graph it. To make a graph, we will first need a list of many values of x and many corresponding values of $\pi(x)$. We do this with two commands. The first might take a minute to compute.

```
In [176]: x_values = numpy.linspace(0,1000000,1001) # The numpy array [0,1000,2000,3000,...,10
          pix_values = numpy.array([primes_upto(x) for x in x_values])  # [FUNCTION(x) for x i
```

We created an array of x-values as before. But the creation of an array of y-values (here, called `pix_values` to stand for $\pi(x)$) probably looks strange. We have done two new things!

1. We have used a list comprehension `[primes_upto(x) for x in x_values]` to create a **list** of y-values.
2. We have used numpy.array(LIST) syntax to convert a Python list into a numpy array.

First, we explain the list comprehension. Instead of pulling out values of a list according to a condition, with `[x for x in LIST if CONDITION]`, we have created a new list based on performing a function each element of a list. The syntax, used above, is `[FUNCTION(x) for x in LIST]`. These two methods of list comprehension can be combined, in fact. The most general syntax for list comprehension is `[FUNCTION(x) for x in LIST if CONDITION]`.

Second, a list comprehension can be carried out on a numpy array, but the result is a plain Python list. It will be better to have a numpy array instead for what follows, so we use the `numpy.array()` function to convert the list into a numpy array.

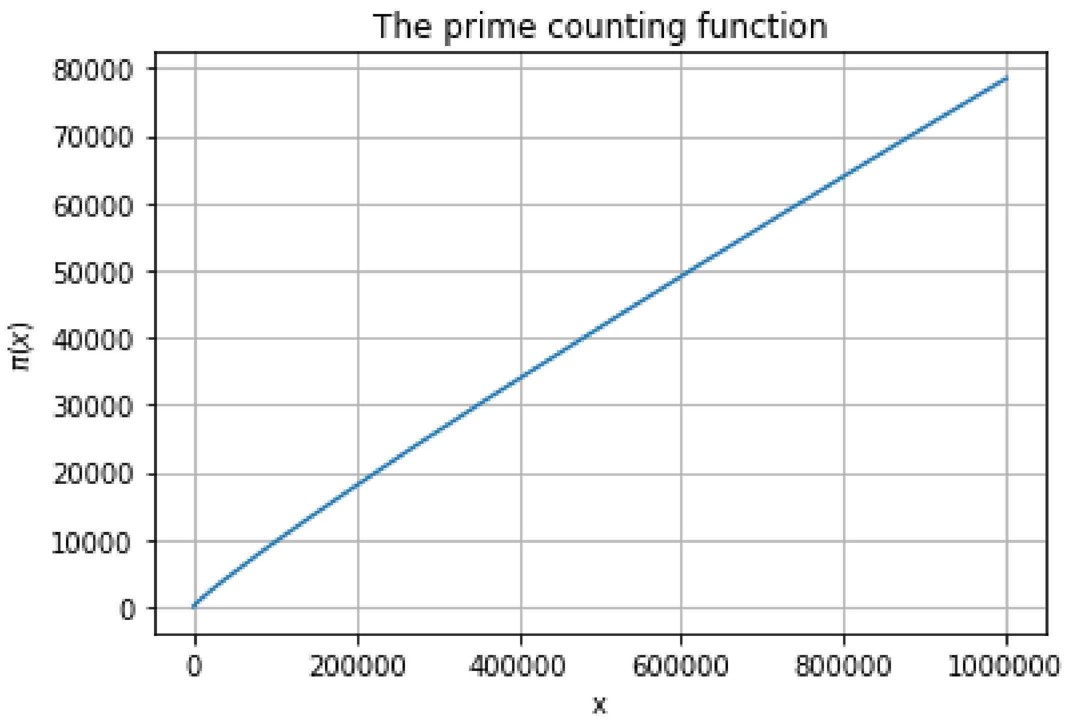
```
In [177]: type(numpy.array([1,2,3]))  # For example.

Out[177]: numpy.ndarray
```

Now we have two numpy arrays: the array of x-values and the array of y-values. We can make a plot with matplotlib.

```
In [178]: len(x_values) == len(pix_values)  # These better be the same, or else matplotlib wil
```

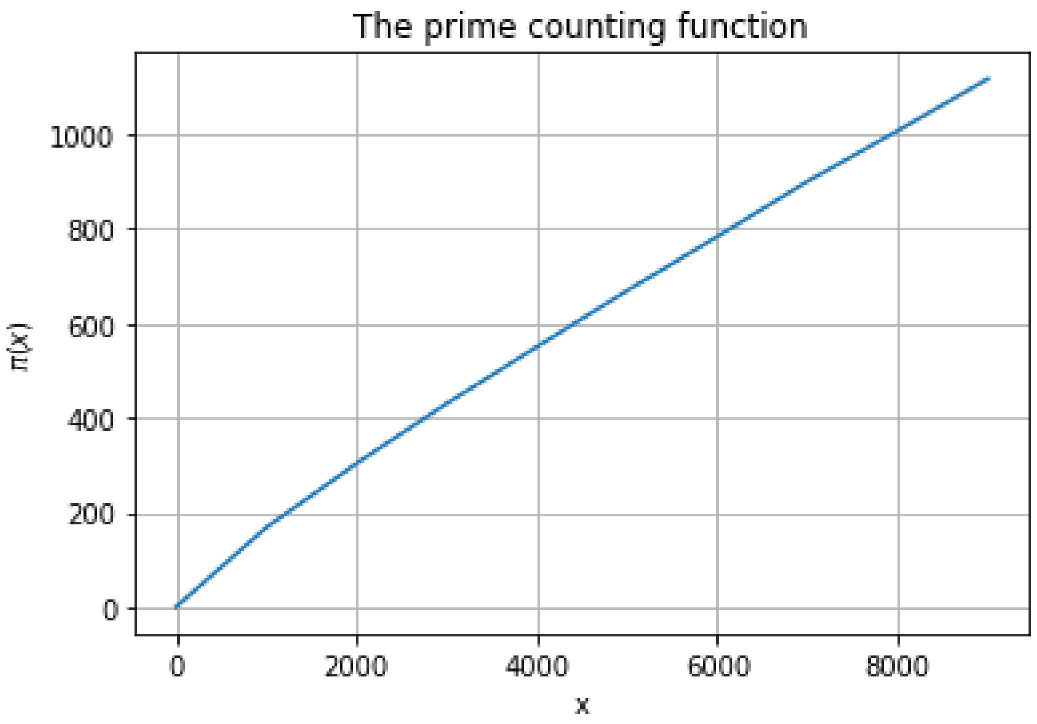
```
Out[178]: True
```

```
In [179]: %matplotlib inline
          plt.plot(x_values, pix_values)
          plt.title('The prime counting function')
          plt.ylabel('$\pi(x)$')
          plt.xlabel('x')
          plt.grid(True)
          plt.show()
```



In this range, the prime counting function might look nearly linear. But if you look closely, there's a subtle downward bend. This is more pronounced in smaller ranges. For example, let's look at the first 10 x-values and y-values only.

```
In [180]: %matplotlib inline
          plt.plot(x_values[:10], pix_values[:10])  # Look closer to 0.
          plt.title('The prime counting function')
          plt.ylabel('$\pi(x)$')
          plt.xlabel('x')
          plt.grid(True)
          plt.show()
```

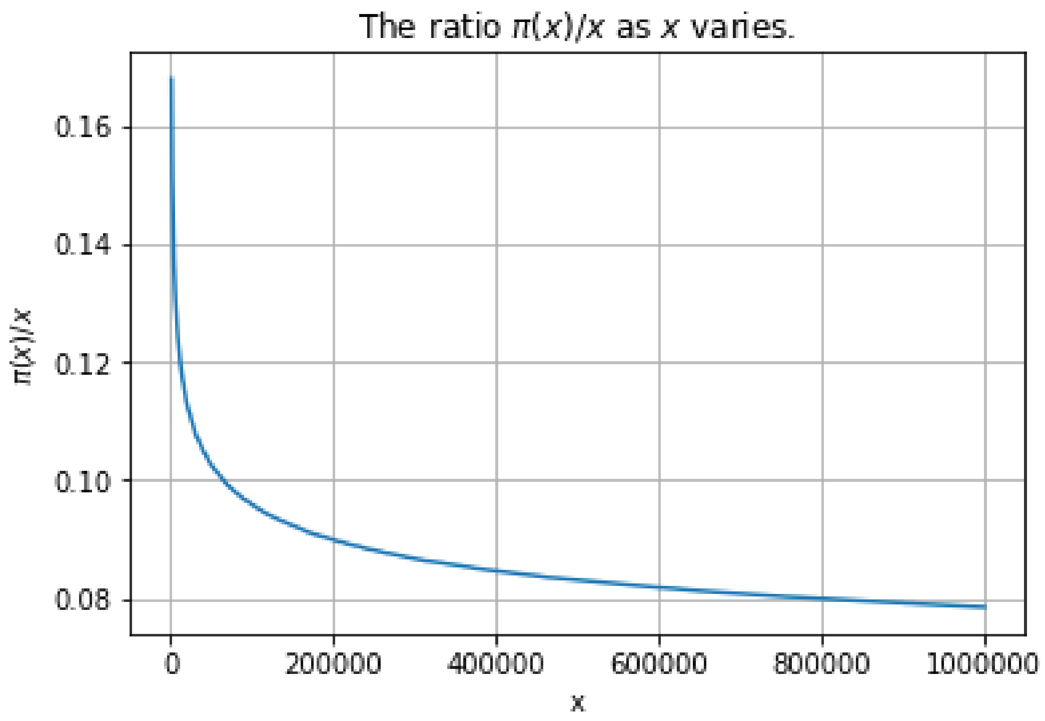


It still looks almost linear, but there's a visible downward bend here. How can we see this bend more clearly? If the graph were linear, its equation would have the form $\pi(x) = mx$ for some fixed slope m (since the graph *does* pass through the origin). Therefore, the quantity $\pi(x)/x$ would be *constant* if the graph were linear.

Hence, if we graph $\pi(x)/x$ on the y-axis and x on the x-axis, and the result is nonconstant, then the function $\pi(x)$ is nonlinear.

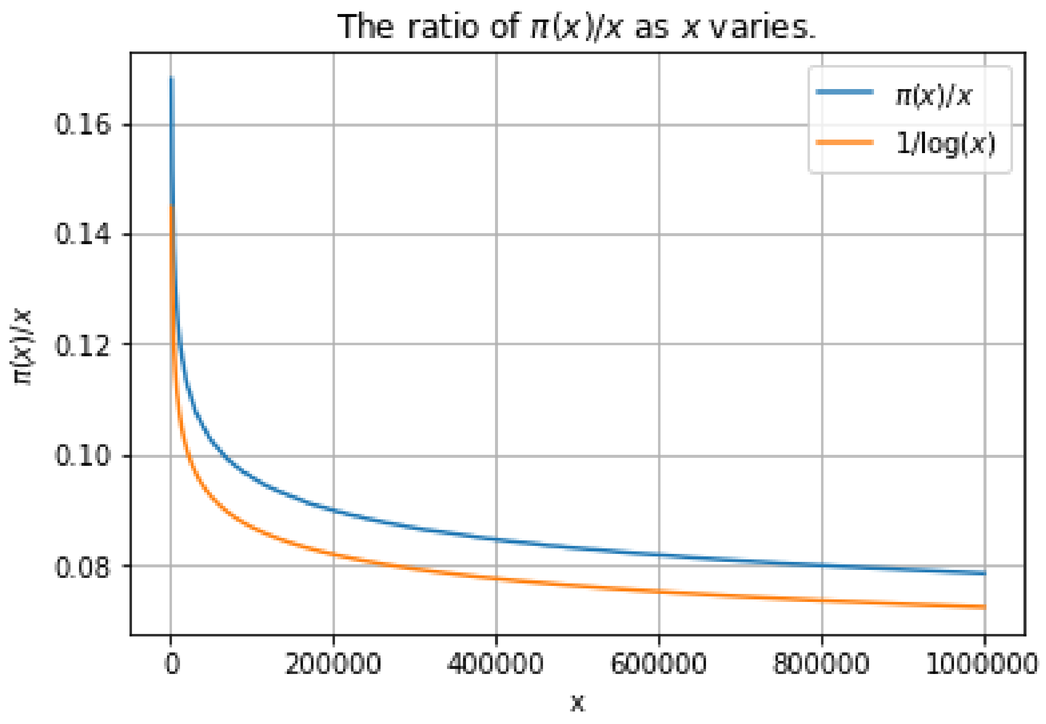
```
In [181]: m_values = pix_values[1:] / x_values[1:]  # We start at 1, to avoid a division by ze
```

```
In [182]: %matplotlib inline
          plt.plot(x_values[1:], m_values)
          plt.title('The ratio $\pi(x) / x$ as $x$ varies.')
          plt.xlabel('x')
          plt.ylabel('$\pi(x) / x$')
          plt.grid(True)
          plt.show()
```



That is certainly not constant! The decay of $\pi(x)/x$ is not so different from $1/\log(x)$, in fact. To see this, let's overlay the graphs. We use the `numpy.log` function, which computes the natural logarithm of its input (and allows an entire array as input).

```
In [183]: %matplotlib inline
          plt.plot(x_values[1:], m_values, label='$\pi(x)/x$')  # The same as the plot above.
          plt.plot(x_values[1:], 1 / numpy.log(x_values[1:]), label='$1 / \log(x)$')  # Overla
          plt.title('The ratio of $\pi(x) / x$ as $x$ varies.')
          plt.xlabel('x')
          plt.ylabel('$\pi(x) / x$')
          plt.grid(True)
          plt.legend()  # Turn on the legend.
          plt.show()
```

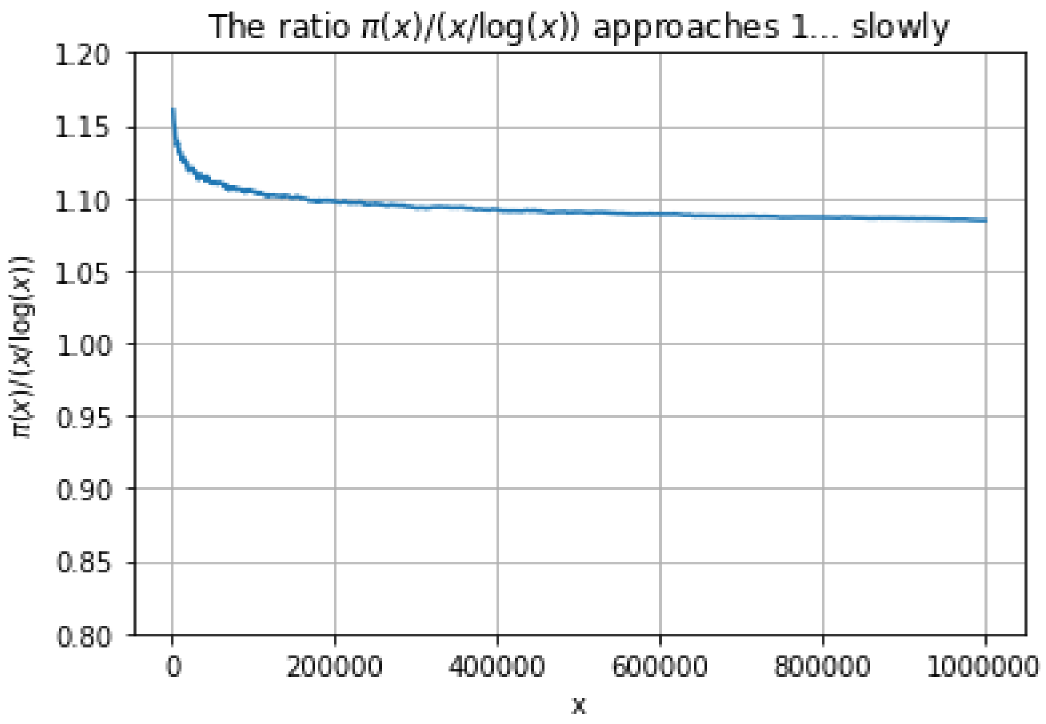


The shape of the decay of $\pi(x)/x$ is very close to $1/\log(x)$, but it looks like there is an offset. In fact, there is, and it is pretty close to $1/\log(x)^2$. And that is close, but again there's another little offset, this time proportional to $2/\log(x)^3$. This goes on forever, if one wishes to approximate $\pi(x)/x$ by an "asymptotic expansion" (not a good idea, it turns out).

The closeness of $\pi(x)/x$ to $1/\log(x)$ is expressed in the **prime number theorem**:

$$\lim_{x \to \infty} \frac{\pi(x)}{x/\log(x)} = 1.$$

```
In [184]: %matplotlib inline
          plt.plot(x_values[1:], m_values * numpy.log(x_values[1:])  )  # Should get closer to
          plt.title('The ratio $\pi(x) / (x / \log(x))$ approaches 1... slowly')
          plt.xlabel('x')
          plt.ylabel('$\pi(x) / (x / \log(x)) $')
          plt.ylim(0.8,1.2)
          plt.grid(True)
          plt.show()
```



Comparing the graph to the theoretical result, we find that the ratio $\pi(x)/(x/\log(x))$ approaches 1 (the theoretical result) but very slowly (see the graph above!).

A much stronger result relates $\pi(x)$ to the "logarithmic integral" $li(x)$. The Riemann hypothesis is equivalent to the statement

$$|\pi(x) - li(x)| = O(\sqrt{x}\log(x)).$$

In other words, the error if one approximates $\pi(x)$ by $li(x)$ is bounded by a constant times $\sqrt{x}\log(x)$. The logarithmic integral function isn't part of Python or numpy, but it is in the mpmath package. If you have this package installed, then you can try the following.

```
In [185]: from mpmath import li
```

```
In [186]: print(primes_upto(1000000))  # The number of primes up to 1 million.
          print(li(1000000))  # The logarithmic integral of 1 million.
```

```
78498
78627.5491594622
```

Not too shabby!

```
In [187]: %timeit primes_upto(1000000)
```

```
4.49 ms ś 164 ţs per loop (mean ś std. dev. of 7 runs, 100 loops each)
```

```
In [188]: %timeit li(1000000)

48.9 ţs ś 424 ns per loop (mean ś std. dev. of 7 runs, 10000 loops each)
```

2.1.3 Prime gaps

As a last bit of data analysis, we consider the **prime gaps**. These are the numbers that occur as differences between consecutive primes. Since all primes except 2 are odd, all prime gaps are even except for the 1-unit gap between 2 and 3. There are many unsolved problems about prime gaps; the most famous might be that a gap of 2 occurs infinitely often (as in the gaps between 3,5 and between 11,13 and between 41,43, etc.).

Once we have our data set of prime numbers, it is not hard to create a data set of prime gaps. Recall that `primes` is our list of prime numbers up to 1 million.

```
In [189]: len(primes) # The number of primes up to 1 million.

Out[189]: 78498

In [190]: primes_allbutlast = primes[:-1]  # This excludes the last prime in the list.
          primes_allbutfirst = primes[1:]  # This excludes the first (i.e., with index 0) prim

In [191]: primegaps = numpy.array(primes_allbutfirst) - numpy.array(primes_allbutlast) # Numpy

In [192]: print(primegaps[:100])  # The first hundred prime gaps!

[ 1  2  2  4  2  4  2  4  6  2  6  4  2  4  6  6  2  6  4  2  6  4  6  8
  4  2  4  2  4 14  4  6  2 10  2  6  6  4  6  6  2 10  2  4  2 12 12  4
  2  4  6  2 10  6  6  6  2  6  4  2 10 14  4  2  4 14  6 10  2  4  6  8
  6  6  4  6  8  4  8 10  2 10  2  6  4  6  8  4  2  4 12  8  4  8  4  6
 12  2 18  6]
```

What have we done? It is useful to try out this method on a short list.

```
In [193]: L = [1,3,7,20]  # A nice short list.

In [194]: print(L[:-1])
          print(L[1:])

[1, 3, 7]
[3, 7, 20]
```

Now we have two lists of the same length. The gaps in the original list `L` are the differences between terms of the *same* index in the two new lists. One might be tempted to just subtract, e.g., with the command `L[1:] - L[:-1]`, but subtraction is not defined for lists.

Fortunately, by converting the lists to numpy arrays, we can use numpy's term-by-term subtraction operation.

```
In [195]: L[1:] - L[:-1]  # This will give a TypeError.  You can't subtract lists!
```

```
        ---------------------------------------------------------------------------

        TypeError                                 Traceback (most recent call last)

        <ipython-input-195-0c27eb74e0a4> in <module>()
    ----> 1 L[1:] - L[:-1]  # This will give a TypeError.  You can't subtract lists!


        TypeError: unsupported operand type(s) for -: 'list' and 'list'
```

```
In [196]: numpy.array(L[1:]) - numpy.array(L[:-1])  # That's better.  See the gaps in the list

Out[196]: array([ 2,  4, 13])
```

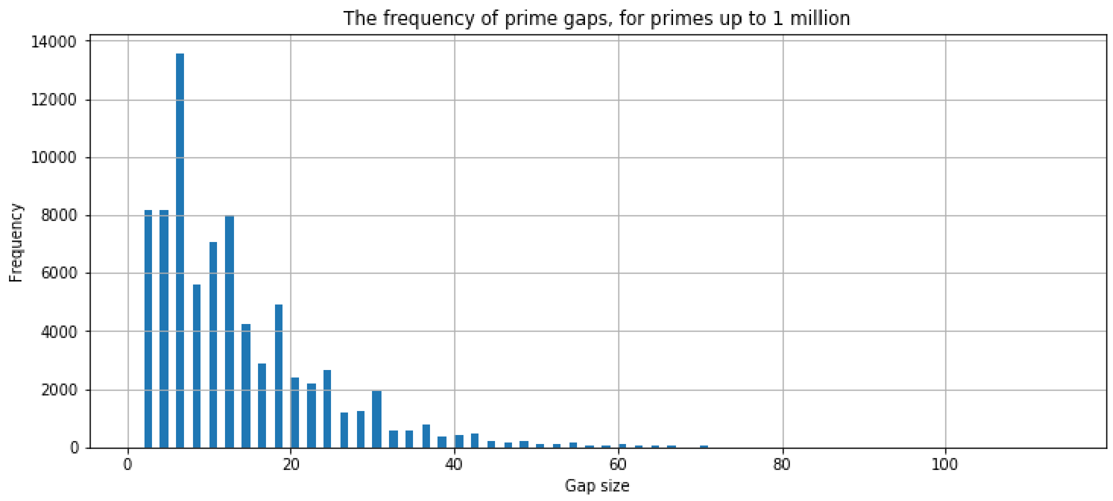
Now let's return to our primegaps data set. It contains all the gap-sizes for primes up to 1 million.

```
In [197]: print(len(primes))
          print(len(primegaps))  # This should be one less than the number of primes.

78498
78497
```

As a last example of data visualization, we use matplotlib to produce a histogram of the prime gaps.

```
In [198]: max(primegaps)  # The largest prime gap that appears!

Out[198]: 114
```

```
In [199]: %matplotlib inline
          plt.figure(figsize=(12, 5))  #  Makes the resulting figure 12in by 5in.
          plt.hist(primegaps, bins=range(1,115)) #  Makes a histogram with one bin for each po
          plt.ylabel('Frequency')
          plt.xlabel('Gap size')
          plt.grid(True)
          plt.title('The frequency of prime gaps, for primes up to 1 million')
          plt.show()
```



Observe that gaps of 2 (twin primes) are pretty frequent. There are over 8000 of them, and about the same number of 4-unit gaps! But gaps of 6 are most frequent in the population, and there are some interesting peaks at 6, 12, 18, 24, 30. What else do you observe?

2.1.4 Exercises

1. Create functions `redprimes_upto(x)` and `blueprimes_upto(x)` which count the number of red/blue primes up to a given number `x`. Recall that we defined red/blue primes to be those of the form 4n+1 or 4n+3, respectively. Graph the relative proportion of red/blue primes as `x` varies from 1 to 1 million. E.g., are the proportions 50%/50% or 70%/30%, and how do these proportions change? Note: this is also visualized in An Illustrated Theory of Numbers and you can read an article by Rubinstein and Sarnak for more.

2. Does there seem to be a bias in the last digits of primes? Note that, except for 2 and 5, every prime ends in 1,3,7, or 9. Note: the last digit of a number `n` is obtained from `n % 10`.

3. Read about the "Prime Conspiracy", recently discovered by Lemke Oliver and Soundararajan. Can you detect their conspiracy in our data set of primes?

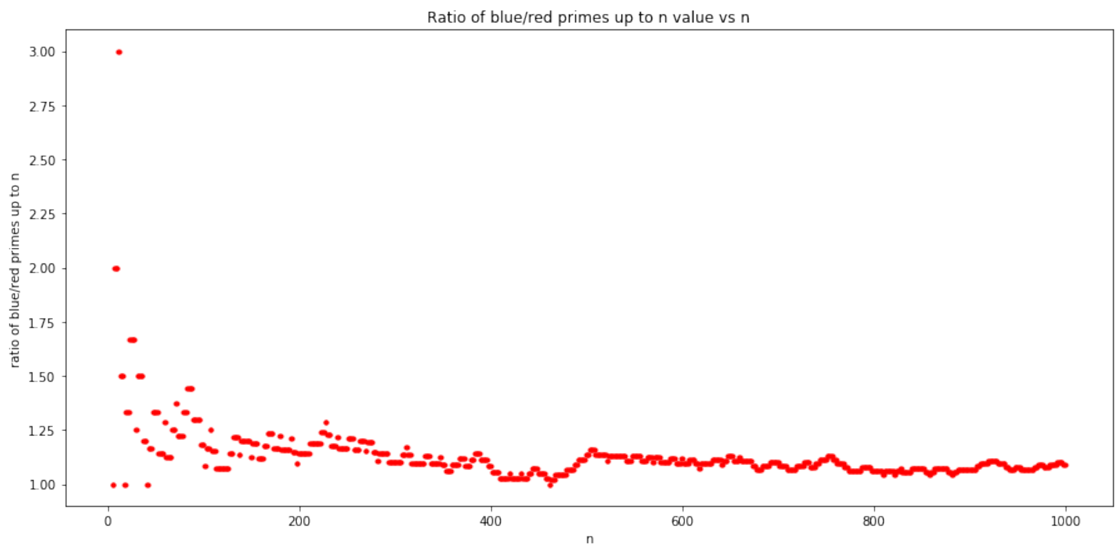
```
In [214]: def redprimes_upto(p):
              primes = where(isprime_list(p))
              return [p for p in primes if p%4 == 1]

In [218]: def blueprimes_upto(p):
              primes = where(isprime_list(p))
              return [p for p in primes if p%4 == 3]

In [361]: redprimes_upto(200)

Out[361]: [5,
           13,
           17,
             29,
             37,
             41,
             53,
             61,
             73,
             89,
             97,
             101,
             109,
             113,
             137,
             149,
             157,
             173,
             181,
             193,
             197]
```

```
In [362]: blueprimes_upto(200)

Out[362]: [3,
             7,
             11,
             19,
             23,
             31,
             43,
             47,
             59,
             67,
             71,
             79,
             83,
             103,
             107,
             127,
             131,
             139,
             151,
             163,
             167,
             179,
             191,
             199]
```

```
In [357]: import matplotlib.pyplot as plt
          def graph_ratio_red_blue_primes(x):
```

```
        # create lists to store how many primes up to x, with x being the index
        reds = []
        blues = []
        ratios = []
        x_axis = []

        for index in range(5,x):
            reds.append(len(redprimes_upto(index)))
            blues.append(len(blueprimes_upto(index)))
            #print(blues[index-1]/reds[index-1])
            ratios.append((blues[index-5]/reds[index-5]))
            x_axis.append(index)
        plt.figure(figsize=(15, 7))
        plt.plot(x_axis, ratios, 'r.')
        plt.xlabel('n')
        plt.ylabel('ratio of blue/red primes up to n')
        plt.title('Ratio of blue/red primes up to n value vs n')
```

```
In [358]: graph_ratio_red_blue_primes(1000)
```

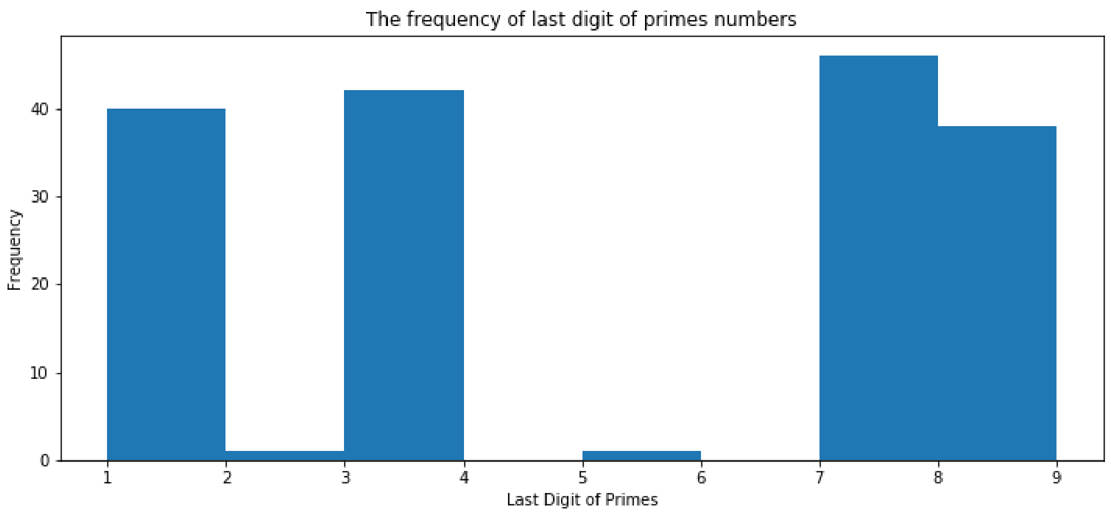


```
In [342]: def last_digit_of_primes_to(x):
              # create array to store last digit of primes
              last_digit = []
              primes = where(isprime_list(x))
              for index in range(len(primes)):
                  last_digit.append(primes[index]%10)
              return last_digit
```

```
In [365]: last_digit_list = last_digit_of_primes_to(500)
          print(last_digit_list)
```

```
[2, 3, 5, 7, 1, 3, 7, 9, 3, 9, 1, 7, 1, 3, 7, 3, 9, 1, 7, 1, 3, 9, 3, 9, 7, 1, 3, 7, 9, 3, 7,
```

```
In [360]: plt.figure(figsize=(12, 5))
          plt.hist(last_digit_list, bins=range(1,10))
          plt.ylabel('Frequency')
          plt.xlabel('Last Digit of Primes')
          #plt.grid(True)
          plt.title('The frequency of last digit of primes numbers')
          plt.show()
```



3 3 I can see the phenomenon being described by printing the last digit of primes up to a given value n (see In [365]), but this can be verified much more concretely by wriging a function to show the ratio of primes ending in 9 followed by ending in 1 vs primes ending in 9 followed by ending in 9 again.