Introduction to Data Science using Python

A Hands-On Guide

Abhijit Dasgupta, Ph.D.

Last updated: June 04, 2020

Contents

1 About this guide

Data science is a broad field that covers data storage, organization, analysis, visualization and reporting. In this guide, we start with a basic primer on the Python language and scientific programming, then progress through the main Python tools for data ingestion, cleaning, manipulation, analyses and presentation. These will primarily consist of the basic usage of the following Python packages:

- pandas
- statsmodels
- scikit-learn
- matplotlib
- seaborn

In my mind this forms the core packages in Python for data analyses and data science work that is applicable to a wide variety of domains.

There are obvious topics that I am not covering here:

- Natural language processing (`nltk`)
- Deep neural networks (`tensorflow`, `keras`, `PyTorch`)
- Image analysis (`scikit-image`, `pillow`)

To my mind these are intermediate to advanced topics, though they are widely used, and so not as foundational to understanding how to use Python for Data Science.

How to use this manual

This manual was originally developed to support a 3 day workshop on using Python for Data Science. Each chapter was covered over roughly half a day, using live coding through the examples.

- **Day 1:** Chapters 2, 3, 4
- **Day 2:** Chapters 5, 6
- **Day 3:** Chapters 7, 8, 9

Each chapter has a fair bit of didactic content describing the methodology underlying the code, to help understand the context for the code. Several datasets were used, which are available in the Github repository for the workshop. Since the workshop, I have discovered that many of the data sets are available online through the Rdatasets package, and so could be loaded directly using `statsmodels`; the examples will slowly be modified accordingly.

2 A Python Primer

2.1 Introduction

Python is a popular, general purpose scripting language. The TIOBE index ranks Python as the third most popular programming language after C and Java, while this recent article in IEEE Computer Society says

> "Python can be used for web and desktop applications, GUI-based desktop applications, machine learning, data science, and network servers. The programming language enjoys immense community support and offers several open-source libraries, frameworks, and modules that make application development a cakewalk." (Belani, 2020)

2.1.1 Python is a modular language

Python is not a monolithic language but is comprised of a base programming language and numerous modules or libraries that add functionality to the language. Several of these libraries are installed with Python. The Anaconda Python Distribution adds more libraries that are useful for data science. Some libraries we will use include `numpy`, `pandas`, `seaborn`, `statsmodels` and `scikit-learn`. In the course of this workshop we will learn how to use Python libraries in your workflow.

2.1.2 Python is a scripting language

Using Python requires typing!! You write *code* in Python that is then interpreted by the Python interpreter to make the computer implement your instructions. **Your code is like a recipe that you write for the computer**. Python is a *high-level language* in that the code is English-like and human-readable and understandable, which reduces the time needed for a person to create the recipe. It is a language in that it has nouns (*variables* or *objects*), verbs (*functions*) and a structure or grammar that allows the programmer to write recipes for different functionalities.

One thing that is important to note in Python: **case is important!**. If we have two objects named `data` and `Data`, they will refer to different things.

Scripting can be frustrating in the beginning. You will find that the code you wrote doesn't work "for some reason", though it looks like you wrote it fine. The first things I look for, in order, are

1. Did I spell all the variables and functions correctly
2. Did I close all the brackets I have opened
3. Did I finish all the quotes I started, and paired single- and double-quotes
4. Did I already import the right module for the function I'm trying to use.

These may not make sense right now, but as we go into Python, I hope you will remember these to help debug your code.

2.2 An example

Let's consider the following piece of Python code:

```
# set a splitting point
split_point = 3

# make two empty lists
lower = []; upper = []

# Split numbers from 0 to 9 into two groups,
# one lower or equal to the split point and
# one higher than the split point

for i in range(10):  # count from 0 to 9
    if i <= split_point:
        lower.append(i)
    else:
        upper.append(i)

print("lower:", lower)
```

```
lower: [0, 1, 2, 3]
```

```
print("upper:", upper)
```

```
upper: [4, 5, 6, 7, 8, 9]
```

First note that any line (or part of a line) starting with # is a **comment** in Python and is ignored by the interpreter. This makes it possible for us to write substantial text to remind us what each piece of our code does

The first piece of code that the Python interpreter actually reads is

```
split_point = 3
```

This takes the number 3 and stores it in the **variable** `split_point`. Variables are just names where some Python object is stored. It really works as an address to some particular part of your computer's memory, telling the Python interpreter to look for the value stored at that particular part of memory. Variable names allow your code to be human-readable since it allows you to write expressive names to remind yourself what you are storing. The rules of variable names are:

1. Variable names must start with a letter or underscore
2. The rest of the name can have letters, numbers or underscores
3. Names are case-sensitive

The next piece of code initializes two **lists**, named `lower` and `upper`.

```
lower = []; upper = []
```

The semi-colon tells Python that, even though written on the same line, a particular instruction ends at the semi-colon, then another piece of instruction is written.

Lists are a catch-all data structure that can store different kinds of things, In this case we'll use them to store numbers.

The next piece of code is a **for-loop** or a loop structure in Python.

```
for i in range(10):  # count from 0 to 9
    if i <= split_point:
        lower.append(i)
    else:
        upper.append(i)
```

It basically works like this:

1. State with the numbers 0-9 (this is achieved in `range(10)`)
2. Loop through each number, naming it `i` each time
    1. Computer programs allow you to over-write a variable with a new value
3. If the number currently stored in `i` is less than or equal to the value of `split_point`, i.e., 3 then add it to the list `lower`. Otherwise add it to the list `upper`

Note the indentation in the code. **This is not by accident**. Python understands the extent of a particular block of code within a for-loop (or within a `if` statement) using the indentations. In this segment there are 3 code blocks:

1. The for-loop as a whole (1st indentation)
2. The `if` statement testing if the number is less than or equal to the split point, telling Python what to do if the test is true
3. The `else` statement stating what to do if the test in the `if` statement is false

Every time a code block starts, the previous line ends in a colon (:). The code block ends when the indentation ends. We'll go into these elements in a bit.

The last bit of code prints out the results

```
print("lower:", lower)
```

```
lower: [0, 1, 2, 3]
```

```
print("upper:", upper)
```

```
upper: [4, 5, 6, 7, 8, 9]
```

The `print` statement adds some text, and then prints out a representation of the object stored in the variable being printed. In this example, this is a list, and is printed as

```
lower: [0, 1, 2, 3]
upper: [4, 5, 6, 7, 8, 9]
```

We will expand on these concepts in the next few sections.

2.2.1 Some general rules on Python syntax

1. Comments are marked by #
2. A statement is terminated by the end of a line, or by a `;`.
3. Indentation specifies blocks of code within particular structures. Whitespace at the beginning of lines matters. Typically you want to have 2 or 4 spaces to specify indentation, not a tab () character. This can be set up in your IDE.
4. Whitespace within lines does not matter, so you can use spaces liberally to make your code more readable
5. Parentheses (`()`) are for grouping pieces of code or for calling functions.

There are several conventions about code styling including the one in PEP8 (PEP = Python Enhancement Proposal) and one proposed by Google. We will typically be using lower case names, with words separated by underscores, in this workshop, basically following PEP8. Other conventions are of course allowed as long as they are within the basic rules stated above.

2.3 Data types in Python

We start with objects in Python. Objects can be of different types, including numbers (integers and floats), strings, arrays (vectors and matrices) and others. Any object can be assigned to a name, so that we can refer to the object in our code. We'll start with the basic types of objects.

2.3.1 Numeric variables

The following is a line of Python code, where we assign the value 1.2 to the variable `a`.

> The act of assigning a name is done using the = sign. This is not equality in the mathematical sense, and has some non-mathematical behavior, as we'll see

```
a = 1.2
```

This is an example of a *floating-point number* or a decimal number, which in Python is called a `float`. We can verify this in Python itself.

```
type(a)
```

```
<class 'float'>
```

Floating point numbers can be entered either as decimals or in scientific notation

```
x = 0.0005
y = 5e-4 # 5 x 10^(-4)
print(x == y)
```

```
True
```

You can also store integers in a variable. Integers are of course numbers, but can be stored more efficiently on your computer. They are stored as an *integer* type, called `int`

```
b = 23
type(b)
```

```
<class 'int'>
```

These are the two primary numerical data types in Python. There are some others that we don't use as often, called `long` (for *long integers*) and `complex` (for *complex numbers*)

2.3.1.1 Operations on numbers

There is an arithmetic and logic available to operate on elemental data types. For example, we do have addition, subtraction , multiplication and division available. For example, for numbers, we can do the following:

Operation	Result
x + y	The sum of x and y
x - y	The difference of x and y
x * y	The product of x and y
x / y	The quotient of x and y
- x	The negative of x
abs(x)	The absolute value of x
x ** y	x raised to the power y
int(x)	Convert a number to integer
float(x)	Convert a number to floating point

Let's see some examples:

```
x = 5
y = 2
```

```
x + y
```

```
7
```

```
x - y
```

```
3
```

```
x * y
```

```
10
```

```
x / y
```

```
2.5
```

```
x ** y
```

```
25
```

2.3.2 Strings

Strings are how text is represented within Python. It is always represented as a quoted object using either single (`''`) or double (`""`) quotes, as long as the types of quotes are matched. For example:

```
first_name = "Abhijit"
last_name = "Dasgupta"
```

The data type that these are stored in is `str`.

```
type(first_name)
```

```
<class 'str'>
```

2.3.2.1 Operations

Strings also have some "arithmetic" associated with it, which involves, essentially, concatenation and repetition. Let's start by considering two character variables that we've initialized.

```
a = "a"
b = "b"
```

Then we get the following operations and results

Operation	Result
a + b	'ab'
a * 5	'aaaaa'

We can also see if a particular character or character string is part of an exemplar string

```
last_name = "Dasgupta"
"gup" in last_name
```

```
True
```

String manipulation is one of the strong suites of Python. There are several *functions* that apply to strings, that we will look at throughout the workshop, and especially when we look at string manipulation. In particular, there are built-in functions in base Python and powerful *regular expression* capabilities in the `re` module.

2.3.3 Truthiness

Truthiness means evaluating the truth of a statement. This typically results in a Boolean object, which can take values `True` and `False`, but Python has several equivalent representations. The following values are considered the same as False:

> `None`, `False`, zero (`0`, `0L`, `0.0`), any empty sequence (`[]`, `''`, `()`), and a few others

All other values are considered True. Usually we'll denote truth by `True` and the number `1`.

2.3.3.1 Operations

We will typically test for the truth of some comparisons. For example, if we have two numbers stored in `x` and `y`, then we can perform the following comparisons

Operation	Result
x < y	x is strictly less than y
x <= y	x is less than or equal to y
x == y	x equals y (note, it's 2 = signs)
x != y	x is not equal to y
x > y	x is strictly greater than y
x >= y	x is greater or equal to y

We can chain these comparisons using Boolean operations

Operation	Result
x | y	Either x is true or y is true or both
x & y	Both x and y are true
not x	if x is true, then false, and vice versa

For example, if we have a number stored in x, and want to find out if it is between 3 and 7, we could write

```
(x >= 3) & (x <= 7)
```

```
True
```

2.3.3.2 A note about variables and types

Some computer languages like C, C++ and Java require you to specify the type of data that will be held in a particular variable. For example,

```
int x = 4;
float y = 3.25;
```

If you try later in the program to assign something of a different type to that variable, you will raise an error. For example, if I did, later in the program, `x = 3.95;`, that would be an error in C.

Python is **dynamically typed**, in that the same variable name can be assigned to different data types in different parts of the program, and the variable will simply be "overwritten". (This is not quite correct. What actually happens is that the variable name now "points" to a different part of the computer's memory where the new data is then stored in appropriate format). So the following is completely fine in Python:

```
x = 4  # An int
x = 3.5  # A float
x = "That's my mother"  # A str
x = True  # A bool
```

> Variables are like individual ingredients in your recipe. It's *mis en place* or setting the table for any operations (*functions*) we want to do to them. In a language context, variables are like *nouns*, which will be acted on by verbs (*functions*). In the next section we'll look at collections of variables. These collections are important in that it allows us to organize our variables with some structure.

2.4 Data structures in Python

Python has several in-built data structures. We'll describe the three most used ones:

1. Lists (`[]`)
2. Tuples (`()`)
3. Dictionaries or dicts (`{}`)

Note that there are three different kinds of brackets being used.

Lists are baskets that can contain different kinds of things. They are ordered, so that there is a first element, and a second element, and a last element, in order. However, the *kinds* of things in a single list doesn't have to be the same type.

Tuples are basically like lists, except that they are *immutable*, i.e., once they are created, individual values can't be changed. They are also ordered, so there is a first element, a second element and so on.

Dictionaries are **unordered** key-value pairs, which are very fast for looking up things. They work almost like hash tables. Dictionaries will be very useful to us as we progress towards the PyData stack. Elements need to be referred to by *key*, not by position.

2.4.1 Lists

```
test_list = ["apple", 3, True, "Harvey", 48205]
test_list
```

```
['apple', 3, True, 'Harvey', 48205]
```

There are various operations we can do on lists. First, we can determine the length (or size) of the list

```
len(test_list)
```

```
5
```

The list is a catch-all, but we're usually interested in extracting elements from the list. This can be done by *position*, since lists are *ordered*. We can extract the 1st element of the list using

```
test_list[0]
```

```
'apple'
```

> Wait!! The index is 0?
>
> Yup. Python is based deep underneath on the C language, where counting starts at 0. So the first element has index 0, second has index 1, and so on. So you need to be careful if you're used to counting from 1, or, if you're used to R, which does start counting at 1.

We can also extract a set of consecutive elements from a list, which is often convenient. The typical form is to write the index as `a:b`. The (somewhat confusing) rule is that `a:b` means that you start at index `a`, but continue until **before index b**. So the notation `2:5` means include elements with index 2, 3, and 4. In the Python world, this is called **slicing**.

```
test_list[2:5]
```

```
[True, 'Harvey', 48205]
```

If you want to start at the beginning or go to the end, there is a shortcut notation. The same rule holds, though. `:3` does **not** include the element at index 3, but `2:` **does** include the element at index 2.

```
test_list[:3]
```

```
['apple', 3, True]
```

```
test_list[2:]
```

```
[True, 'Harvey', 48205]
```

The important thing here is if you provide an index `a:b`, then `a` is include but `b` **is not**.

You can also count **backwards** from the end. The last element in a Python list has index `-1`.

index	0	1	2	3	4
element	'apple'	3	True	'Harvey'	48205
counting backwards	-5	-4	-3	-2	-1

```
test_list[-1]
```

```
48205
```

You can also use negative indices to denote sequences within the list, with the same indexing rule applying. Note that you count from the last element (-1) and go backwards.

```
test_list[:-1]
```

```
['apple', 3, True, 'Harvey']
```

```
test_list[-3:]
```

```
[True, 'Harvey', 48205]
```

```
test_list[-3:-1]
```

```
[True, 'Harvey']
```

You can also make a list of lists, or nested lists

```
test_nested_list = [[1, "a", 2, "b"], [3, "c", 4, "d"]]
test_nested_list
```

```
[[1, 'a', 2, 'b'], [3, 'c', 4, 'd']]
```

This will come in useful when we talk about arrays and data frames.

You can also check if something is in the list, i.e. is a member.

```
"Harvey" in test_list
```

```
True
```

Lists have the following properties

- They can be heterogenous (each element can be a different type)
- Lists can hold complex objects (lists, dicts, other objects) in addition to atomic objects (single numbers or words)
- List have an ordering, so you can access list elements by position
- List access can be done counting from the beginning or the end, and consecutive elements can be extracted using slices.

2.4.2 Tuples

Tuples are like lists, except that once you create them, you can't change them. This is why tuples are great if you want to store fixed parameters or entities within your Python code, since they can't be over-written even by mistake. You can extract elements of a tuple, but you can't over-write them. This is called *immutable*.

Note that, like lists, tuples can be heterogenous, which is also useful for coding purposes, as we will see.

```
test_tuple = ("apple", 3, True, "Harvey", 48205)
```

```
test_tuple[:3]
```

```
('apple', 3, True)
```

```
test_list[0] = "pear"
test_list
```

```
['pear', 3, True, 'Harvey', 48205]
```

See what happens in the next bit of code

```
test_tuple[0] = "pear"
test_tuple
```

(I'm not running this since it gives an error)

> Tuples are like lists, but once created, they cannot be changed. They are ordered and can be sliced.

2.4.3 Dictionaries

Dictionaries, or `dict`, are collections of key-value pairs. Each element is referred to by *key*, not by *index*. In a dictionary, the keys can be strings, numbers or tuples, but the values can be any Python object. So you could have a dictionary where one value is a string, another is a number and a third is a DataFrame (essentially a data set, using the pandas library). A simple example might be an entry in a list of contacts

```
contact = {
    "first_name": "Abhijit",
    "last_name": "Dasgupta",
    "Age": 48,
    "address": "124 Main St",
    "Employed": True,
}
```

Note the special syntax. You separate the key-value pairs by colons (`:`), and each key-value pair is separated by commas. If you get a syntax error creating a dict, look at these first.

If you try to get the first name out using an index, you run into an error:

```
contact[0]
```

```
Error in py_call_impl(callable, dots$args, dots$keywords): KeyError: 0
```

You need to extract it by key

```
contact["first_name"]
```

```
'Abhijit'
```

A dictionary is mutable, so you can change the value of any particular element

```
contact["address"] = "123 Main St"
contact["Employed"] = False
contact
```

```
{'first_name': 'Abhijit', 'last_name': 'Dasgupta', 'Age': 48, 'address': '123 Main St', 'Em
```

You can see all the keys and values in a dictionary using extractor functions

```
contact.keys()
```

```
dict_keys(['first_name', 'last_name', 'Age', 'address', 'Employed'])
```

```
contact.values()
```

```
dict_values(['Abhijit', 'Dasgupta', 48, '123 Main St', False])
```

It turns out that dictionaries are really fast in terms of retrieving information, without having to count where an element it. So it is quite useful

We'll see that dictionaries are also one way to easily create pandas DataFrame objects on the fly.

There are a couple of other ways to create dict objects. One is using a list of tuples. Each key-value pair is represented by a tuple of length 2, where the 1st element is the key and the second element is the value.

```
A = [('first_name','Abhijit'),('last_name','Dasgupta'),('address', '124 Main St')]
dict(A)
```

```
{'first_name': 'Abhijit', 'last_name': 'Dasgupta', 'address': '124 Main St'}
```

This actually can be utilized to create a dict from a pair of lists. There is a really neat function, `zip`, that inputs several lists of the same length and creates a list of tuples, where the i-th element of each tuple comes from the i-th list, in order.

```
A = ['first_name', 'last_name','address']
B = ['Abhijit','Dasgupta','124 Main St']
dict(zip(A, B))
```

```
{'first_name': 'Abhijit', 'last_name': 'Dasgupta', 'address': '124 Main St'}
```

> The `zip` function is quite powerful in putting several lists together with corresponding elements of each list into a tuple

On a side note, there is a function `defaultdict` from the `collections` module that is probably better to use. We'll come back to it when we talk about modules.

2.5 Operational structures in Python

2.5.1 Loops and list comprehensions

Loops are a basic construct in computer programming. The basic idea is that you have a recipe that you want to repeatedly run on different entities that you have created. The crude option would be to copy and paste your code several times, changing whatever inputs change across the entities. This is not only error-prone, but inefficient given that loops are a standard element of all programming languages.

You can create a list of these entities, and, using a loop, run your recipe on each entity automatically. For example, you have a data about votes in the presidential election from all 50 states, and you want to figure out what the percent voting for each major party is. So you could write this recipe in pseudocode as

```
Start with a list of datasets, one for each state
for each state
    compute and store fraction of votes that are Republican
    compute and store fraction of votes that are Democratic
```

This is just English, but it can be translated easily into actual code. We'll attempt that at the end of this section.

The basic idea of a list is that there is a list of things you want to iterate over. You create a dummy variable as stand-in for each element of that list. Then you create a for-loop. This works like a conveyor belt and basket, so to speak. You line up elements of the list on the conveyor belt, and as you run the loop, one element of the list is "scooped up" and processed. Once that processing is done, the next element is "scooped up", and so forth. The dummy variable is essentially the basket (so the same basket (variable name) is re-used over and over until the conveyor belt (list) is empty).

In the examples below, we are showing a common use of for loops where we are enumerating the elements of a list as 0, 1, 2, ... using `range(len(test_list))`. So the dummy variable `i` takes values 0, 1, 2, ... until the length of the list is reached. For each value of `i`, this for loop prints the i-th element of `test_list`.

```
for i in range(len(test_list)):
    print(test_list[i])
```

```
pear
3
True
Harvey
48205
```

Sometimes using the index number is easier to understand. However, we don't need to do this. We can just send the list itself into the for-loop (u) now is the dummy variable containing the actual element of `test_list`. We'll get the same answer.

```
for u in test_list:
    print(u)
```

```
pear
3
True
Harvey
48205
```

The general structure for a `for` loop is:

```
for (element) in (list):
      do some stuff
      do more stuff
```

As a more practical example, let's try and sum a set of numbers using a for-loop (we'll see much better ways of doing this later)

```
test_list2 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
mysum = 0
for u in test_list2:
    mysum = mysum + u
print(mysum)
```

```
55
```

There are two things to note here.

1. The code `mysum = mysum + u` is perfectly valid, once you realize that this isn't really math but an assignment or pointer to a location in memory. This code says that you find the current value stored in `mysum`, add the value of `u` to it, and then store it back into the storage that `mysum` points to
2. Indentation matters! Indent the last line and see what happens when you run this code

2.5.1.1 A little deeper

The entity to the right of the `in` in the for-loop can be an **iterator**, which is a generalization of a list. For example, we used `range(len(test_list2))` above. If we just type

```
range(10)
```

```
range(0, 10)
```

nothing really happens. This is an example of an iterator, which is only evaluated when it is called, rather than being stored in memory. This is useful especially when you iterate over large numbers of things, in terms of preserving memory and speed. To see the corresponding list, you would do

```
list(range(10))
```

```
[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
```

This range iterator is quite flexible:

```
list(range(5, 10))  # range from 5 to 10
```

```
[5, 6, 7, 8, 9]
```

```
list(range(0, 10, 2))  # range from 0 to 10 by 2
```

```
[0, 2, 4, 6, 8]
```

Note the rules here are very much like the slicing rules.

Other iterators that are often useful are the `enumerate` iterator and the `zip` iterator.

`enumerate` automatically creates both the index and the value for each element of a list.

```
L = [0, 2, 4, 6, 8]
for i, val in enumerate(L):
    print(i, val)
```

```
0 0
1 2
2 4
3 6
4 8
```

`zip` puts multiple lists together and creates a composite iterator. You can have any number of iterators in zip, and the length of the result is determined by the length of the shortest iterator. We introduced an example of `zip` as a way to create a `dict`.

Technically, `zip` can take multiple *iterators* as inputs, not just lists

```
first = ["Han", "Luke", "Leia", "Anakin"]
last = ["Solo", "Skywalker", "Skywaker", "Skywalker"]
types = ['light','light','light','light/dark/light']

for val1, val2, val3 in zip(first, last, types):
    print(val1, val2, ' : ', val3)
```

```
Han Solo  :  light
Luke Skywalker  :  light
Leia Skywaker  :  light
Anakin Skywalker  :  light/dark/light
```

2.5.1.2 Controlling loops

There are two statements that can affect how loops run:

- The `break` statement breaks out of the loop
- The `continue` statement skips the rest of the current loop and continues to the next element

For example

```
x = list(range(10))

for u in x:
    if u % 2 == 1: # If u / 2 gives a remainder of 1
        continue
    if u >= 8:
        break
    print(u)
```

```
0
2
4
6
```

In this loop, we don't print the odd numbers, and we stop the loop once it gets to 8.

2.5.2 List comprehensions

List comprehensions are quick ways of generating a list from another list by using some recipe. For example, if we wanted to create a list of the squares of all the numbers in `test_list2`, we could write

```
squares = [u ** 2 for u in test_list2]
squares
```

```
[1, 4, 9, 16, 25, 36, 49, 64, 81, 100]
```

Similarly, if we wanted to find out what the types of each element of `test_tuple` is, we could use

```
[type(u) for u in test_tuple]
```

```
[<class 'str'>, <class 'int'>, <class 'bool'>, <class 'str'>, <class 'int'>]
```

Exercise: Can you use a list comprehension to find out the types of each element of the `contact` dict?

We can also use list comprehensions to extract arbitrary sets of elements of lists

```
test = ['a','b','c','d','e','f','g']
test1 = [test[i] for i in [0,2,3,5]]
test1
```

```
['a', 'c', 'd', 'f']
```

2.5.3 Conditional evaluations

The basic structure for conditional evaluation of code is an **if-then-else** structure.

```
if Condition 1 is true then
    do Recipe 1
else if (elif) Condition 2 is true then
  do Recipe 2
else
  do Recipe 3
```

In Python, this is implemented as a `if-elif-else` structure. Let's take an example where we have a list of numbers, and we want to record whether the number is negative, odd, or even.

```
x = [-2, -1, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
y = []  # an empty list

for u in x:
    if u < 0:
        y.append("Negative")
    elif u % 2 == 1:  # what is remainder when dividing by 2
        y.append("Odd")
    else:
        y.append("Even")

print(y)
```

```
['Negative', 'Negative', 'Even', 'Odd', 'Even', 'Odd', 'Even', 'Odd', 'Even', 'Odd', 'Even'
```

Note here that the indentation (leading whitespace) is crucial to this structure. The `if-elif-else` structure is embedded in a for-loop, so the entire structure in indented. Also, each particular recipe is also indented within the if-elif-else structure.

> The `elif` is optional, in that if you have only 2 conditions, then an `if-else` structure is sufficient. However, you can have multiple `elif`'s if you have more conditions. This kind of structure has to start with an `if`, end with an `else` and can have 0 or more `elif` in the middle.

2.6 Functions

We've already seen some examples of **functions**, such as the `print()` function. For example, if we write `print(y)`, the function name is `print` and the functions *argument* is `y`. So what are functions?

Functions are basically encapsulated recipes. They are groups of code that are given a name and can be called with 0 or more arguments. In a cookbook, you might have a recipe for pasta primavera. This is the name of a recipe that has ingredients and a method to cook. In Python, a similar recipe for the mean might be as follows:

```
def my_mean(x):
    y = 0
    for u in x:
        y += u
    y = y / len(x)
    return y
```

This takes a list of numbers `x`, loops over the elements of `x` to find their sum, and then divides by the length of `x` to compute the mean. It then returns this mean.

> The notation `+=` is a shortcut often used in programming. The statement `y += u` means, take the current value of `y`, add the value of `u` to it, and store it back in to `y`. This is a shorthand for `y = y + u`. In analogous fashion, you can use `-=`, `*=` and `/=` to do subtraction, multiplication and division respectively.

A Python function must start with the keyword `def` followed by the name of the function, the arguments within parentheses, and then a colon. The actual code for the function is indented, just like in for-loops and if-elif-else structures. It ends with a `return` function which specifies the output of the function.

To use the `my_mean` function,

```
x = list(range(10))
my_mean(x)
```

```
4.5
```

2.6.1 Documenting your functions

Python has an in-built documentation system that allows you to readily document your functions using *docstrings*. Basically, right after the first line with `def`, you can create a (multi-line) string that documents the function and will be printed if the help system is used for that function. You can create a multi-line string by **bounding it with 3 quotation marks on each side**. For example,

```
def my_mean(x):
    """
  A function to compute the mean of a list of numbers.

  INPUTS:
  x : a list containing numbers

  OUTPUT:
  The arithmetic mean of the list of numbers
  """
    y = 0
    for u in x:
        y = y + u
    y = y / len(x)
    return y
```

```
help(my_mean)
```

```
Help on function my_mean in module __main__:

my_mean(x)
    A function to compute the mean of a list of numbers.

    INPUTS:
    x : a list containing numbers

    OUTPUT:
    The arithmetic mean of the list of numbers
```

2.7 Modules and Packages

Python itself was built with the principle "Batteries included", in that it already comes with useful tools for a wide variety of tasks. On top of that, there is a large ecosystem of third-party tools and packages that can be added on to add more functionality. Almost all the data science functionality in Python comes from third-party packages.

2.7.1 Using modules

The Python standard library as well as third-party packages (which I'll use interchangeably with the term libraries) are structured as modules. In order to use a particular module you have to "activate" it in your Python session using the `import` statement.

```
import math

math.cos(math.pi)
```

```
-1.0
```

In these statements, we have imported the `math` module. This module has many functions, one of which is the cosine or `cos` function. We use the notation `math.cos` to let Python know that we want to use the `cos` function that is in the `math` module. The value of π is also stored in the `math` module as `math.pi`, ie. the element `pi` within the moduel `math`.

Modules can often have long names, so Python caters to our laziness by allowing us to create aliases for modules when we import them. In this workshop we will use the following statements quite often

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

These statements import 3 modules into the current Python session, namely `numpy`, `pandas` and a sub-module of the `matplotlib` module called `pyplot`. In each case, we have provided an alias to the module that is imported. So, in subsequent calls, we can just use the aliases.

```
np.cos(np.pi)
```

```
-1.0
```

If we only want some particular components of a module to be imported, we can specify them using the `from ... import ...` syntax. These imported components will not need the module specification when we subsequently use them.

```
from math import pi, sin, cos

print(sin(pi))
```

```
1.2246467991473532e-16
```

```
print(cos(pi))
```

```
-1.0
```

We had made reference to the `defaultdict` function from the `collections` module before. Using this instead of `dict` can be advantagous sometimes in data scientific work.

```
from collections import defaultdict

A = defaultdict(list) # Specify each component will be a list
B = {}

s = [('yellow', 1), ('blue', 2), ('yellow', 3), ('blue', 4), ('red', 1)]
for k,v in s: # k = key, v = value
    B[k].append(v)
```

```
Error in py_call_impl(callable, dots$args, dots$keywords): KeyError: 'yellow'
```

```
for k,v in s:
    A[k].append(v)
A
```

```
defaultdict(<class 'list'>, {'yellow': [1, 3], 'blue': [2, 4], 'red': [1]})
```

The `defaultdict` sees a new key, and adds it to the dict, initializing it with an empty list (since we specified `defaultdict(list)`. The normal dict requires the key to already be in place in the dict for any operations to take place on that key-value pair. So the `default` dict is safer for on-the-fly work and when we don't know beforehand what keys we will encounter when storing data into the dict.

> There is a temptation to use this method to import everything in a module so you don't have to specify the module. This is a **bad practice** generally, both because you clutter up the namespace that Python reads from, and because you may unknowingly over-write and replace a function from one module with one from another module, and you will have a hard time debugging your code.
>
> The code you do **NOT** want to use is
>
> ```
> from math import *
> ```

2.7.2 Useful modules in Python's standard library

Module	Description
`os` and `sys`	Interfacing with the operating system, including files, directories, and executing shell commands
`math` and `cmath`	Mathematical functions
`itertools`	Constructing and using iterators
`random`	Generate random numbers
`collections`	More general collections for objects, beyond lists, tuples and dicts

2.7.3 Installing third-party packages/libraries

The Anaconda Python distribution comes with its own installer and package manager called `conda`. The Anaconda repository contains most of the useful packages for data science, and many come pre-installed with the distribution. However, you can easily install packages using the `conda` manager.

```
conda install pandas
```

would install the `pandas` package into your Python installation. If you wanted a particular version of this package, you could use

```
conda install pandas=0.23
```

to install version 0.23 of the `pandas` package.

Anaconda also provides a repository for user-created packages. For example, to install the Python package RISE which I use for creating slides from Jupyter notebooks, I use

```
conda install -c conda-forge rise
```

Sometimes you may find a Python package that is not part of the Anaconda repositories. Then you can use the more general Python program `pip` to install packages

```
pip install supersmoother
```

This goes looking in the general Python package repository PyPi, which you can also search on a web browser.

2.8 Environments

One of the nice things about Python is that you can set up environments for particular projects, that have all the packages you need for that project, without having to install those packages system-wide. This practice is highly recommended, since it creates a sandbox for you to play in for a project without contaminating the code from another project.

The Anaconda distribution and the `conda` program make this quite easy. There are a couple of ways of doing this.

2.8.1 Command-line/shell

You can open up a command line terminal (any terminal on Mac and Linux, the Anaconda Terminal in Windows) to create a new environment. For example, I have an environment I call **ds** that is my data science environment. This will include the packages `numpy`, `scipy`, `pandas`,`matplotlib`, `seaborn`,`statsmodels` and `scikit-learn` in it. The quick way to do this is

```
conda create -n ds numpy scipy pandas matplotlib seaborn statsmodels scikit-
learn
```

To use this environment, at the command line, type

```
conda activate ds
```

Once you're done using it, at the command line, type

```
conda deactivate
```

When your environment is activated, you'll see the name of the environment before the command prompt

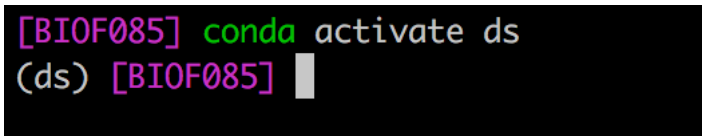


Figure 2.1: image-20200511003754816

2.8.2 Using Anaconda Navigator

Open the Anaconda Navigator from your start menu or using Spotlight on a Mac. Within the app is a section named "Environments"

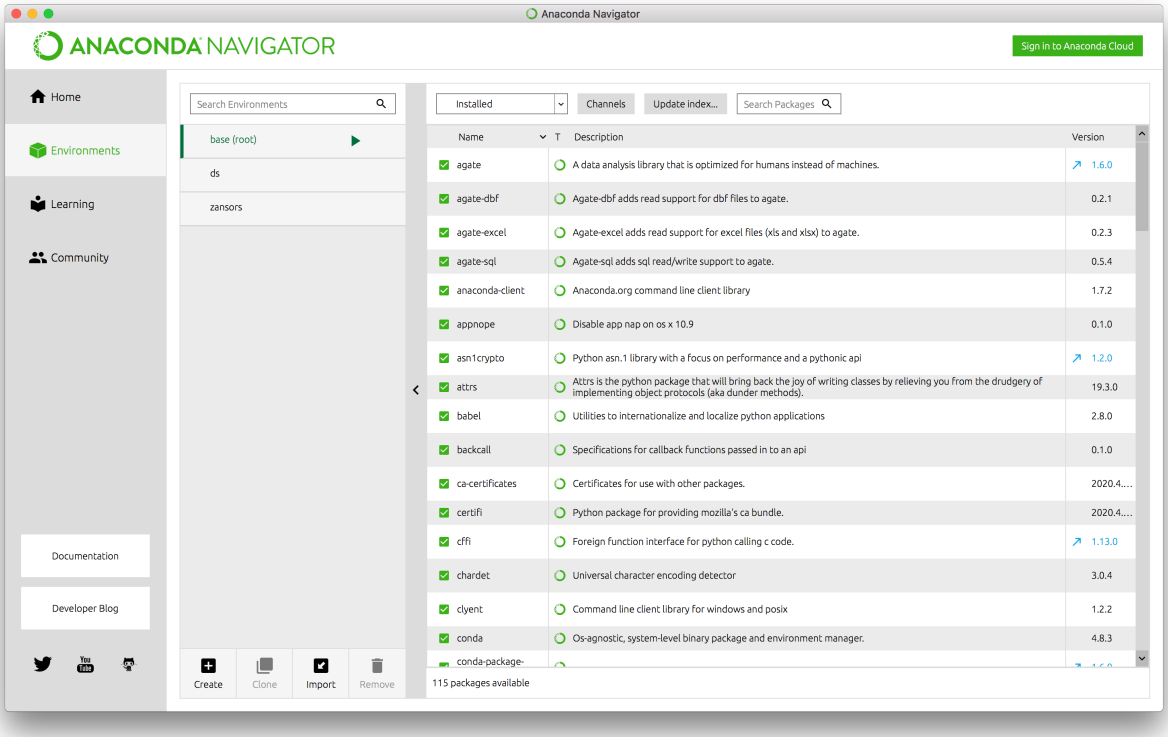


Figure 2.2: image-20200511004048781

At the bottom of the 2nd pane, you can see a "Create" button. Clicking it creates a pop-up window.

I've named this new environment `ds1` since I already have a `ds` environment. Click "Create". You'll have to wait a bit of time for the environment to be created. You can then add/install packages to this environment by clicking on packages on the right panel, making sure you changed the first drop-down menu from "Installed" to "Not installed".

Once you've selected the packages you want to install, click "Appy" on the bottom right of the window.

To activate an environment, you can go to the Home pane for Anaconda Navigator and change the environment on the "Applications on" drop-down menu.

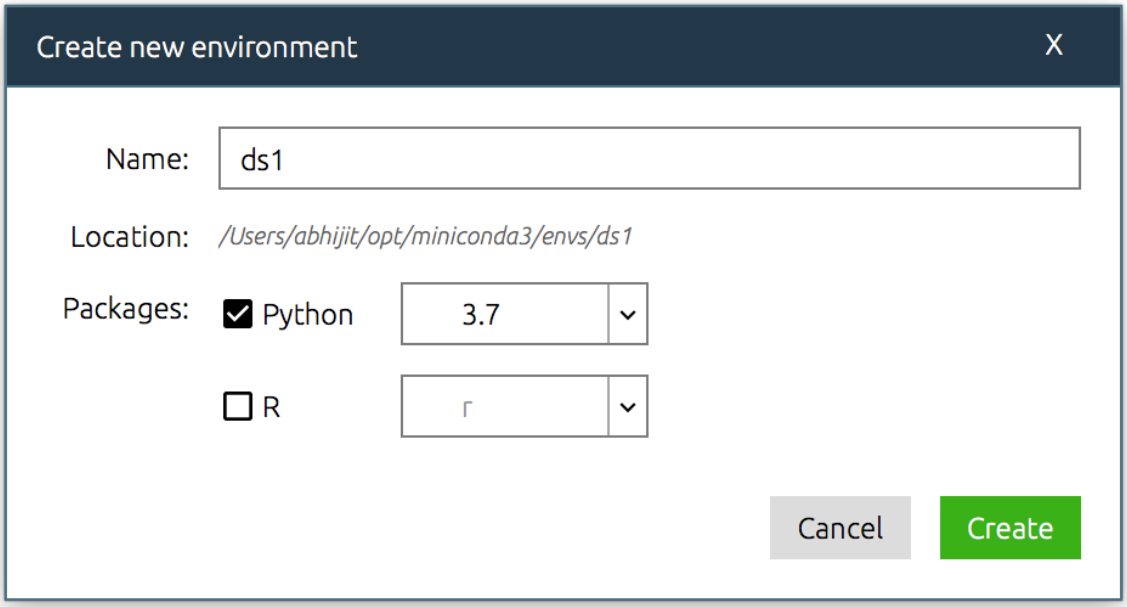


Figure 2.3: image-20200511004259459

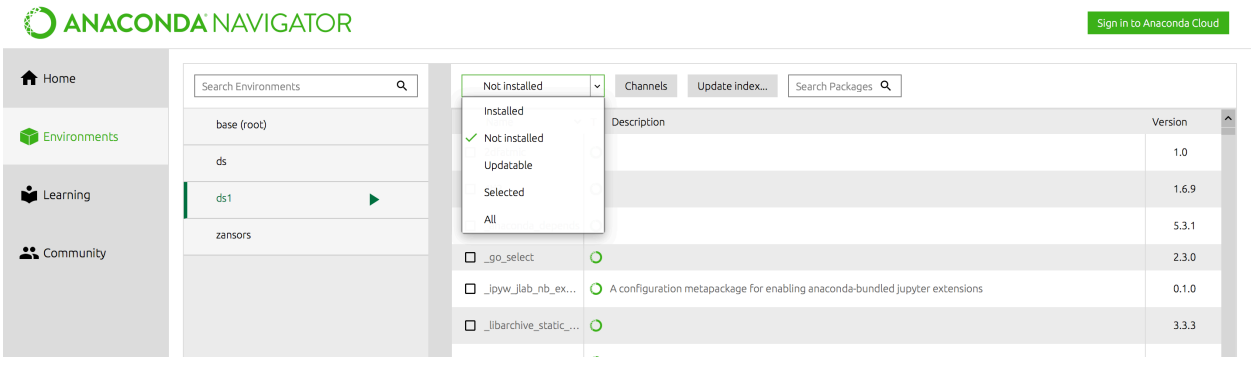


Figure 2.4: image-20200511004530076

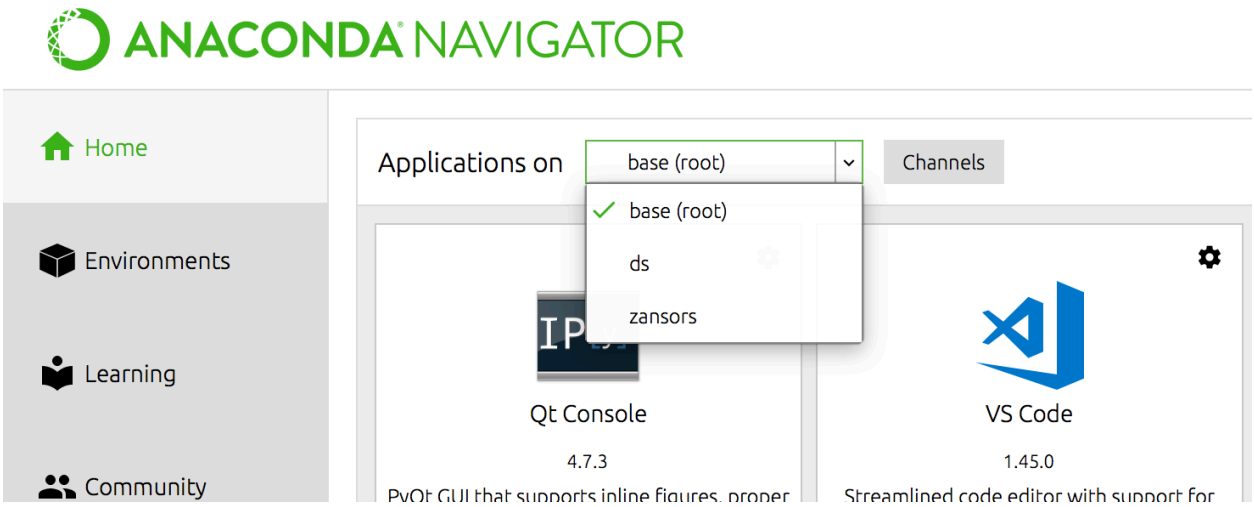


Figure 2.5: image-20200511004920105

2.8.3 Reproducing environments

Suppose you've got an environment set up the way you like it, and want to clone it on another machine that has Anaconda installed. There is an easy way to do this. You have to use the command line (Anaconda Prompt (Win) or a terminal) for this.

First activate the environment you want to export (I'll use `ds` as an example)

```
conda activate ds
```

Then export the environment specifications which includes all the packages installed in that environment

```
conda env export > environment.yml
```

You can take this `environment.yml` file to a new computer, or e-mail it to a collaborator to install the environment. This environment can be created on the new computer using

```
conda env create -f environment.yml
```

where the first line of the `environment.yml` file creates the environment name.

You can also create the environment from an `environment.yml` file from Anaconda Navigator by using the Import button rather than the Create button in the instructions above.

> If you are changing operating systems, create the `environment.yml` file using the command
>
> ```
> conda env export --from-history > environment.yml
> ```
>
> This avoids potential issues with dependencies that may not be compatible across operating systems

2.9 Seeking help

Most Python functions have some amount of documentation. As we saw when we created our own function, this documentation is part of the function definition. It can be accessed at the Python console in 2 ways:

```
help(sum)
```

```
Help on built-in function sum in module builtins:

sum(iterable, /, start=0)
    Return the sum of a 'start' value (default: 0) plus an iterable of numbers

    When the iterable is empty, return the start value.
    This function is intended specifically for use with numeric values and may
    reject non-numeric types.
```

or

```
# sum?
```

You can see the documentation of the `my_sum` function we created earlier in this way, as well.

Other resources that are your friends in the internet age are

1. Stack Overflow: This is a Q & A site. To find Python-related questions, use the tag `python`.
2. Google: Of course.
3. Cross-Validated: A data science oriented Q & A site. Once again, use the tag `python`.

3 Python tools for data science

(last updated 2020-05-18)

3.1 The PyData Stack

The Python Data Stack comprises a set of packages that makes Python a powerful data science language. These include

- Numpy: provides arrays and matrix algebra
- Scipy: provides scientific computing capabilities
- matplotlib: provides graphing capabilities

These were the original stack that was meant to replace Matlab. However, these were meant to tackle purely numerical data, and the kinds of heterogeneous data we regularly face needed more tools. These were added more recently.

- Pandas: provides data analytic structures like the data frame, as well as basic descriptive statistical capabilities
- statsmodels: provides a fairly comprehensive set of statistical functions
- scikit-learn: provides machine learning capabilities

This is the basic stack of packages we will be using in this workshop. Additionally we will use a few packages that add some functionality to the data science process. These include

- seaborn: Better statistical graphs
- plotly: Interactive graphics
- biopython: Python for bioinformatics

We may also introduce the package `rpy2` which allows one to run R from within Python. This can be useful since many bioinformatic pipelines are already implemented in R.

> The PyData stack also includes `sympy`, a symbolic mathematics package emulating Maple

3.2 Numpy (numerical and scientific computing)

We start by importing the Numpy package into Python using the alias `np`.

```
import numpy as np
```

Numpy provides both arrays (vectors, matrices, higher dimensional arrays) and vectorized functions which are very fast. Let's see how this works.

```
z = [1,2,3,4,5,6,7,8,9.3,10.6] # This is a list
z_array = np.array(z)
z_array
```

```
array([ 1. ,  2. ,  3. ,  4. ,  5. ,  6. ,  7. ,  8. ,  9.3, 10.6])
```

Now, we have already seen functions in Python earlier. In Numpy, there are functions that are optimized for arrays, that can be accessed directly from the array objects. This is an example of *object-oriented programming* in Python, where functions are provided for particular *classes* of objects, and which can be directly accessed from the objects. We will use several such functions over the course of this workshop, but we won't actually talk about how to do this program development here.

> Numpy functions are often very fast, and are *vectorized*, i.e., they are written to work on vectors of numbers rather than single numbers. This is an advantage in data science since we often want to do the same operation to all elements of a column of data, which is essentially a vector

We apply the functions `sum`, `min` (minimum value) and `max` (maximum value) to `z_array`.

```
z_array.sum()
```

```
55.9
```

```
z_array.min()
```

```
1.0
```

```
z_array.max()
```

```
10.6
```

The versions of these functions in Numpy are optimized for arrays and are quite a bit faster than the corresponding functions available in base Python. When doing data work, these are the preferred functions.

These functions can also be used in the usual function manner:

```
np.max(z_array)
```

```
10.6
```

Calling `np.max` ensures that we are using the `max` function from numpy, and not the one in base Python.

3.2.1 Numpy data types

Numpy arrays are homogeneous in type.

```
np.array(['a','b','c'])
```

```
array(['a', 'b', 'c'], dtype='<U1')
```

```
np.array([1,2,3,6,8,29])
```

```
array([ 1,  2,  3,  6,  8, 29])
```

But, what if we provide a heterogeneous list?

```
y = [1,3,'a']
np.array(y)
```

```
array(['1', '3', 'a'], dtype='<U21')
```

So what's going on here? Upon conversion from a heterogeneous list, numpy converted the numbers into strings. This is necessary since, by definition, numpy arrays can hold data of a single type. When one of the elements is a string, numpy casts all the other entities into strings as well. Think about what would happen if the opposite rule was used. The string 'a' doesn't have a corresponding number, while both numbers 1 and 3 have corresponding string representations, so going from string to numeric would create all sorts of problems.

> The advantage of numpy arrays is that the data is stored in a contiguous section of memory, and you can be very efficient with homogeneous arrays in terms of manipulating them, applying functions, etc. However, numpy does provide a "catch-all" `dtype` called `object`, which can be any Python object. This `dtype` essentially is an array of pointers to actual data stored in different parts of the memory. You can get to the actual objects by extracting them. So one could do

```
np.array([1,3,'a'], dtype='object')
```

```
array([1, 3, 'a'], dtype=object)
```

> which would basically be a valid numpy array, but would go back to the actual objects when used, much like a list. We can see this later if we want to transform a heterogeneous `pandas` `DataFrame` into a numpy array. It's not particularly useful as is, but it prevents errors from popping up during transformations from `pandas` to numpy.

3.2.2 Generating data in numpy

We had seen earlier how we could generate a sequence of numbers in a list using `range`. In numpy, you can generate a sequence of numbers in an array using `arange` (which actually creates the array rather than provide an iterator like `range`).

```
np.arange(10)
```

```
array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
```

You can also generate regularly spaced sequences of numbers between particular values

```
np.linspace(start=0, stop=1, num=11) # or np.linspace(0, 1, 11)
```

```
array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ])
```

You can also do this with real numbers rather than integers.

```
np.linspace(start = 0, stop = 2*np.pi, num = 10)
```

```
array([0.        , 0.6981317 , 1.3962634 , 2.0943951 , 2.7925268 ,
       3.4906585 , 4.1887902 , 4.88692191, 5.58505361, 6.28318531])
```

More generally, you can transform lists into numpy arrays. We saw this above for vectors. For matrices, you can provide a list of lists. Note the double `[` in front and back.

```
np.array([[1,3,5,6],[4,3,9,7]])
```

```
array([[1, 3, 5, 6],
       [4, 3, 9, 7]])
```

You can generate an array of 0's

```
np.zeros(10)
```

```
array([0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])
```

This can easily be extended to a two-dimensional array (a matrix), by specifying the dimension of the matrix as a tuple.

```
np.zeros((10,10))
```

```
array([[0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0., 0., 0., 0., 0., 0.]])
```

You can also generate a matrix of 1s in a similar manner.

```
np.ones((3,4))
```

```
array([[1., 1., 1., 1.],
       [1., 1., 1., 1.],
       [1., 1., 1., 1.]])
```

In matrix algebra, the identity matrix is important. It is a square matrix with 1's on the diagonal and 0's everywhere else.

```
np.eye(4)
```

```
array([[1., 0., 0., 0.],
       [0., 1., 0., 0.],
       [0., 0., 1., 0.],
       [0., 0., 0., 1.]])
```

You can also create numpy vectors directly from lists, as long as lists are made up of atomic elements of the same type. This means a list of numbers or a list of strings. The elements can't be more composite structures, generally. One exception is a list of lists, where all the lists contain the same type of atomic data, which, as we will see, can be used to create a matrix or 2-dimensional array.

```
a = [1,2,3,4,5,6,7,8]
b = ['a','b','c','d','3']

np.array(a)
```

```
array([1, 2, 3, 4, 5, 6, 7, 8])
```

```
np.array(b)
```

```
array(['a', 'b', 'c', 'd', '3'], dtype='<U1')
```

3.2.2.1 Random numbers

Generating random numbers is quite useful in many areas of data science. All computers don't produce truly random numbers but generate *pseudo-random* sequences. These are completely deterministic sequences defined algorithmically that emulate the properties of random numbers. Since these are deterministic, we can set a *seed* or starting value for the sequence, so that we can exactly reproduce this sequence to help debug our code. To actually see how things behave in simulations we will often run several sequences of random numbers starting at different seed values.

The seed is set by the `RandomState` function within the `random` submodule of numpy. Note that all Python names are case-sensitive.

```
rng = np.random.RandomState(35) # set seed
rng.randint(0, 10, (3,4))
```

```
array([[9, 7, 1, 0],
       [9, 8, 8, 8],
       [9, 7, 7, 8]])
```

We have created a 3x4 matrix of random integers between 0 and 10 (in line with slicing rules, this includes 0 but not 10).

We can also create a random sample of numbers between 0 and 1.

```
rng.random_sample((5,2))
```

```
array([[0.04580216, 0.91259827],
       [0.21381599, 0.3036373 ],
       [0.98906362, 0.1858815 ],
       [0.98872484, 0.75008423],
       [0.22238605, 0.14790391]])
```

We'll see later how to generate random numbers from particular probability distributions.

3.2.3 Vectors and matrices

Numpy generates arrays, which can be of arbitrary dimension. However the most useful are vectors (1-d arrays) and matrices (2-d arrays).

In these examples, we will generate samples from the Normal (Gaussian) distribution, with mean 0 and variance 1.

```
A = rng.normal(0,1,(4,5))
```

We can compute some characteristics of this matrix's dimensions. The number of rows and columns are given by `shape`.

```
A.shape
```

```
(4, 5)
```

The total number of elements are given by `size`.

```
A.size
```

```
20
```

If we want to create a matrix of 0's with the same dimensions as `A`, we don't actually have to compute its dimensions. We can use the `zeros_like` function to figure it out.

```
np.zeros_like(A)
```

```
array([[0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0.],
       [0., 0., 0., 0., 0.]])
```

We can also create vectors by only providing the number of rows to the random sampling function. The number of columns will be assumed to be 1.

```
B = rng.normal(0, 1, (4,))
B
```

```
array([-0.45495378,  1.04307172,  0.70451207, -0.6171649 ])
```

3.2.3.1 Extracting elements from arrays

The syntax for extracting elements from arrays is almost exactly the same as for lists, with the same rules for slices.

Exercise: State what elements of `B` are extracted by each of the following statements

```
B[:3]
B[:-1]
B[[0,2,4]]
B[[0,2,5]]
```

For matrices, we have two dimensions, so you can slice by rows, or columns or both.

```
A
```

```
array([[-2.45677354,  0.36686697, -0.20453263, -0.54380446,  0.09524207],
       [ 1.06236144,  1.03937554,  0.01247733, -0.35427727, -1.18997812],
       [ 0.95554288,  0.30781478,  0.7328766 , -1.28670314, -1.03870027],
       [-0.81398211, -1.02506031,  0.12407205,  1.21491023, -1.44645123]])
```

We can extract the first column by specifying `:` (meaning everything) for the rows, and the index for the column (reminder, Python starts counting at 0)

```
A[:,0]
```

```
array([-2.45677354,  1.06236144,  0.95554288, -0.81398211])
```

Similarly the 4th row can be extracted by putting the row index, and `:` for the column index.

```
A[3,:]
```

```
array([-0.81398211, -1.02506031,  0.12407205,  1.21491023, -1.44645123])
```

All slicing operations work for rows and columns

```
A[:2,:2]
```

```
array([[-2.45677354,  0.36686697],
       [ 1.06236144,  1.03937554]])
```

3.2.3.2 Array operations

We can do a variety of vector and matrix operations in numpy.

First, all usual arithmetic operations work on arrays, like adding or multiplying an array with a scalar.

```
A = rng.randn(3,5)
A
```

```
array([[-0.15367796,  2.50215522,  0.19420725,  0.54928294, -1.1737166 ],
       [ 1.11456557,  0.07447758,  1.58518354,  1.61986225, -0.24616333],
       [-0.02682273,  0.2196577 ,  0.41680351, -0.86319929,  0.50355595]])
```

```
A + 10
```

```
array([[ 9.84632204, 12.50215522, 10.19420725, 10.54928294,  8.8262834 ],
       [11.11456557, 10.07447758, 11.58518354, 11.61986225,  9.75383667],
       [ 9.97317727, 10.2196577 , 10.41680351,  9.13680071, 10.50355595]])
```

We can also add and multiply arrays **element-wise** as long as they are the same shape.

```
B = rng.randint(0,10, (3,5))
B
```

```
array([[6, 2, 3, 9, 8],
       [5, 9, 3, 9, 7],
       [0, 4, 2, 5, 0]])
```

```
A + B
```

```
array([[ 5.84632204,  4.50215522,  3.19420725,  9.54928294,  6.8262834 ],
       [ 6.11456557,  9.07447758,  4.58518354, 10.61986225,  6.75383667],
       [-0.02682273,  4.2196577 ,  2.41680351,  4.13680071,  0.50355595]])
```

```
A * B
```

```
array([[-0.92206775,  5.00431043,  0.58262175,  4.94354649, -9.38973278],
       [ 5.57282784,  0.67029821,  4.75555061, 14.57876027, -1.72314331],
       [-0.        ,  0.8786308 ,  0.83360701, -4.31599644,  0.        ]])
```

You can also do **matrix multiplication**. Recall what this is.

If you have a matrix A_{mxn} and another matrix B_{nxp}, as long as the number of columns of A and rows of B are the same, you can multiply them ($C_{mxp} = A_{mxn}B_{nxp}$), with the (i,j)-th element of C being

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}, i = 1, \dots, m; j = 1, \dots, p$$

In numpy the operant for matrix multiplication is `@`.

In the above examples, `A` and `B` cannot be multiplied since they have incompatible dimensions. However, we can take the *transpose* of `B`, i.e. flip the rows and columns, to make it compatible with `A` for matrix multiplication.

```
A @ np.transpose(B)
```

```
array([[ 0.21867814, 19.0611592 , 13.14345008],
       [24.20135281, 23.85429363, 11.56758865],
       [-2.21155643, -1.15068575, -2.60375863]])
```

```
np.transpose(A) @ B
```

```
array([[  4.65076009,   9.61644327,   2.82901737,   8.51387483,
          6.57253531],
       [ 15.38531919,   6.55323945,   8.16921379,  24.28798365,
         20.53858478],
       [  9.09116118,  16.32228035,   6.17177937,  18.0985346 ,
         12.64994275],
       [ 11.39500892,  12.22452901,   4.78103701,  15.20631032,
         15.73329932],
       [ -8.27311623,  -2.54867934,  -3.25252787, -10.26113957,
        -11.11287609]])
```

More generally, you can *reshape* a numpy array into a new shape, provided it is compatible with the number of elements in the original array.

```
D = rng.randint(0,5, (4,4))
D
```

```
array([[0, 2, 0, 0],
       [4, 0, 0, 4],
       [0, 3, 2, 0],
       [3, 0, 0, 3]])
```

```
D.reshape(8,2)
```

```
array([[0, 2],
       [0, 0],
       [4, 0],
       [0, 4],
       [0, 3],
       [2, 0],
       [3, 0],
       [0, 3]])
```

```
D.reshape(1,16)
```

```
array([[0, 2, 0, 0, 4, 0, 0, 4, 0, 3, 2, 0, 3, 0, 0, 3]])
```

This can also be used to cast a vector into a matrix.

```
e = np.arange(20)
E = e.reshape(5,4)
E
```

```
array([[ 0,  1,  2,  3],
       [ 4,  5,  6,  7],
       [ 8,  9, 10, 11],
       [12, 13, 14, 15],
       [16, 17, 18, 19]])
```

One thing to note in all the reshaping operations above is that the new array takes elements of the old array **by row**. See the examples above to convince yourself of that.

3.2.3.3 Statistical operations on arrays

You can sum all the elements of a matrix using `sum`. You can also sum along rows or along columns by adding an argument to the `sum` function.

```
A = rng.normal(0, 1, (4,2))
A
```

```
array([[-1.60798054, -0.05162306],
       [-0.49218049, -0.1262316 ],
       [ 0.56927597,  0.05438786],
       [ 0.33120322, -0.81820729]])
```

```
A.sum()
```

```
-2.141355938226197
```

You can sum along rows (i.e., down columns) with the option `axis = 0`

```
A.sum(axis=0)
```

```
array([-1.19968185, -0.94167409])
```

You can sum along columns (i.e., across rows) with `axis = 1`.

```
A.sum(axis=1)
```

```
array([-1.6596036 , -0.61841209,  0.62366383, -0.48700407])
```

Of course, you can use the usual function calls: `np.sum(A, axis = 1)`

We can also find the minimum and maximum values.

```
A.min(axis = 0)
```

```
array([-1.60798054, -0.81820729])
```

```
A.max(axis = 0)
```

```
array([0.56927597, 0.05438786])
```

We can also find the **position** where the minimum and maximum values occur.

```
A.argmin(axis=0)
```

```
array([0, 3])
```

```
A.argmax(axis=0)
```

```
array([2, 2])
```

We can sort arrays and also find the indices which will result in the sorted array. I'll demonstrate this for a vector, where it is more relevant

```
a = rng.randint(0,10, 8)
a
```

```
array([9, 2, 6, 6, 4, 4, 3, 4])
```

```
np.sort(a)
```

```
array([2, 3, 4, 4, 4, 6, 6, 9])
```

```
np.argsort(a)
```

```
array([1, 6, 4, 5, 7, 2, 3, 0])
```

```
a[np.argsort(a)]
```

```
array([2, 3, 4, 4, 4, 6, 6, 9])
```

`np.argsort` can also help you find the 2nd smallest or 3rd largest value in an array, too.

```
ind_2nd_smallest = np.argsort(a)[1]
a[ind_2nd_smallest]
```

```
3
```

```
ind_3rd_largest = np.argsort(a)[-3]
a[ind_3rd_largest]
```

```
6
```

You can also sort strings in this way.

```
m = np.array(['Aram','Raymond','Elizabeth','Donald','Harold'])
np.sort(m)
```

```
array(['Aram', 'Donald', 'Elizabeth', 'Harold', 'Raymond'], dtype='<U9')
```

If you want to sort arrays **in place**, you can use the `sort` function in a different way.

```
m.sort()
m
```

```
array(['Aram', 'Donald', 'Elizabeth', 'Harold', 'Raymond'], dtype='<U9')
```

3.2.3.4 Putting arrays together

We can put arrays together by row or column, provided the corresponding axes have compatible lengths.

```
A = rng.randint(0,5, (3,5))
B = rng.randint(0,5, (3,5))
print('A = ', A)
```

```
A =  [[3 4 2 1 3]
 [0 3 1 1 1]
 [4 0 2 0 4]]
```

```
print('B = ', B)
```

```
B =  [[1 4 2 1 3]
 [2 0 3 2 0]
 [4 0 2 3 3]]
```

```
np.hstack((A,B))
```

```
array([[3, 4, 2, 1, 3, 1, 4, 2, 1, 3],
       [0, 3, 1, 1, 1, 2, 0, 3, 2, 0],
       [4, 0, 2, 0, 4, 4, 0, 2, 3, 3]])
```

```
np.vstack((A,B))
```

```
array([[3, 4, 2, 1, 3],
       [0, 3, 1, 1, 1],
       [4, 0, 2, 0, 4],
       [1, 4, 2, 1, 3],
       [2, 0, 3, 2, 0],
       [4, 0, 2, 3, 3]])
```

Note that both `hstack` and `vstack` take a **tuple** of arrays as input.

3.2.3.5 Logical/Boolean operations

You can query a matrix to see which elements meet some criterion. In this example, we'll see which elements are negative.

```
A < 0
```

```
array([[False, False, False, False, False],
       [False, False, False, False, False],
       [False, False, False, False, False]])
```

This is called **masking**, and is useful in many contexts.

We can extract all the negative elements of A using

```
A[A<0]
```

```
array([], dtype=int64)
```

This forms a 1-d array. You can also count the number of elements that meet the criterion

```
np.sum(A<0)
```

```
0
```

Since the entity `A<0` is a matrix as well, we can do row-wise and column-wise operations as well.

3.2.4 Beware of copies

One has to be a bit careful with copying objects in Python. By default, if you just assign one object to a new name, it does a *shallow copy*, which means that both names point to the same memory. So if you change something in the original, it also changes in the new copy.

```
A[0,:]
```

```
array([3, 4, 2, 1, 3])
```

```
A1 = A
A1[0,0] = 4
A[0,0]
```

```
4
```

To actually create a copy that is not linked back to the original, you have to make a *deep copy*, which creates a new space in memory and a new pointer, and copies the original object to the new memory location

```
A1 = A.copy()
A1[0,0] = 6
A[0,0]
```

```
4
```

You can also replace sub-matrices of a matrix with new data, provided that the dimensions are compatible. (Make sure that the sub-matrix we are replacing below truly has 2 rows and 2 columns, which is what `np.eye(2)` will produce)

```
A[:2,:2] = np.eye(2)
A
```

```
array([[1, 0, 2, 1, 3],
       [0, 1, 1, 1, 1],
       [4, 0, 2, 0, 4]])
```

3.2.4.1 Reducing matrix dimensions

Sometimes the output of some operation ends up being a matrix of one column or one row. We can reduce it to become a vector. There are two functions that can do that, `flatten` and `ravel`.

```
A = rng.randint(0,5, (5,1))
A
```

```
array([[2],
       [1],
       [4],
       [2],
       [4]])
```

```
A.flatten()
```

```
array([2, 1, 4, 2, 4])
```

```
A.ravel()
```

```
array([2, 1, 4, 2, 4])
```

So why two functions? I'm not sure, but they do different things behind the scenes. `flatten` creates a **copy**, i.e. a new array disconnected from `A`. `ravel` creates a **view**, so a representation of the original array. If you then changed a value after a `ravel` operation, you would also change it in the original array; if you did this after a `flatten` operation, you would not.

3.2.5 Broadcasting in Python

Python deals with arrays in an interesting way, in terms of matching up dimensions of arrays for arithmetic operations. There are 3 rules:

1. If two arrays differ in the number of dimensions, the shape of the smaller array is padded with 1s on its *left* side
2. If the shape doesn't match in any dimension, the array with shape = 1 in that dimension is stretched to match the others' shape
3. If in any dimension the sizes disagree and none of the sizes are 1, then an error is generated

```
A = rng.normal(0,1,(4,5))
B = rng.normal(0,1,5)
```

```
A.shape
```

```
(4, 5)
```

```
B.shape
```

```
(5,)
```

```
A - B
```

```
array([[ 0.25410957,  1.89009891, -0.87300221, -0.17852271, -0.64735645],
       [ 0.72991872,  4.58821268, -0.03379553, -1.67352398,  1.43345867],
       [-1.51421293,  1.69993003,  1.81140727, -1.71622014,  0.52276992],
       [-1.30611819,  3.29767231,  0.91060221,  0.29490453,  1.24919619]])
```

B is 1-d, A is 2-d, so B's shape is made into (1,5) (added to the left). Then it is repeated into 4 rows to make it's shape (4,5), then the operation is performed. This means that we subtract the first element of B from the first column of A, the second element of B from the second column of A, and so on.

You can be explicit about adding dimensions for broadcasting by using `np.newaxis`.

```
B[np.newaxis,:].shape
```

```
(1, 5)
```

```
B[:,np.newaxis].shape
```

```
(5, 1)
```

3.2.5.1 An example (optional, intermediate/advanced))

This can be very useful, since these operations are faster than for loops. For example:

```
d = rng.random_sample((10,2))
d
```

```
array([[0.9497432 , 0.22672332],
       [0.96319737, 0.61011348],
       [0.2542308 , 0.60550727],
       [0.64054935, 0.85273037],
       [0.19747218, 0.45957414],
       [0.41571736, 0.9902779 ],
       [0.33720945, 0.30637872],
       [0.15139082, 0.30126537],
       [0.72158605, 0.3560211 ],
       [0.86288412, 0.66608767]])
```

We want to find the Euclidean distance (the sum of squared differences) between the points defined by the rows. This should result in a 10x10 distance matrix

```
d.shape
```

```
(10, 2)
```

```
d[np.newaxis,:,:]
```

```
array([[[0.9497432 , 0.22672332],
        [0.96319737, 0.61011348],
        [0.2542308 , 0.60550727],
        [0.64054935, 0.85273037],
        [0.19747218, 0.45957414],
        [0.41571736, 0.9902779 ],
        [0.33720945, 0.30637872],
        [0.15139082, 0.30126537],
        [0.72158605, 0.3560211 ],
        [0.86288412, 0.66608767]]])
```

creates a 3-d array with the first dimension being of length 1

```
d[np.newaxis,:,:].shape
```

```
(1, 10, 2)
```

```
d[:, np.newaxis,:]
```

```
array([[[0.9497432 , 0.22672332]],

       [[0.96319737, 0.61011348]],

       [[0.2542308 , 0.60550727]],

       [[0.64054935, 0.85273037]],

       [[0.19747218, 0.45957414]],

       [[0.41571736, 0.9902779 ]],

       [[0.33720945, 0.30637872]],

       [[0.15139082, 0.30126537]],

       [[0.72158605, 0.3560211 ]],

       [[0.86288412, 0.66608767]]])
```

creates a 3-d array with the 2nd dimension being of length 1

```
d[:,np.newaxis,:].shape
```

```
(10, 1, 2)
```

Now for the trick, using broadcasting of arrays. These two arrays are incompatible without broadcasting, but with broadcasting, the right things get repeated to make things compatible

```
dist_sq = np.sum((d[:,np.newaxis,:] - d[np.newaxis,:,:]) ** 2)
```

```
dist_sq.shape
```

```
()
```

```
dist_sq
```

```
29.512804540441067
```

Whoops! we wanted a 10x10 matrix, not a scalar.

```
(d[:,np.newaxis,:] - d[np.newaxis,:,:]).shape
```

```
(10, 10, 2)
```

What we really want is the 10x10 distance matrix.

```
dist_sq = np.sum((d[:,np.newaxis,:] - d[np.newaxis,:,:]) ** 2, axis=2)
```

You can verify what is happening by creating `D = d[:,np.newaxis,:]-d[np.newaxis,:,:]` and then looking at `D[:,:,0]` and `D[:,:,1]`. These are the difference between each combination in the first and second columns of d, respectively. Squaring and summing along the 3rd axis then gives the sum of squared differences.

```
dist_sq
```

```
array([[0.        , 0.14716903, 0.62721478, 0.48748566, 0.6201312 ,
        0.8681992 , 0.38154258, 0.64292305, 0.0687736 , 0.20058553],
       [0.14716903, 0.        , 0.5026548 , 0.16296469, 0.60899716,
        0.44425934, 0.48411568, 0.75441703, 0.12293897, 0.01319586],
       [0.62721478, 0.5026548 , 0.        , 0.21036128, 0.02451802,
        0.17412634, 0.09636334, 0.1031392 , 0.28066428, 0.37412884],
       [0.48748566, 0.16296469, 0.21036128, 0.        , 0.35088921,
```

```
 0.06946875, 0.39051522, 0.54338972, 0.25328705, 0.08426824],
[0.6201312 , 0.60899716, 0.02451802, 0.35088921, 0.        ,
 0.32927744, 0.04299534, 0.02718516, 0.28541859, 0.48542089],
[0.8681992 , 0.44425934, 0.17412634, 0.06946875, 0.32927744,
 0.        , 0.47388157, 0.54460678, 0.49583735, 0.30505741],
[0.38154258, 0.48411568, 0.09636334, 0.39051522, 0.04299534,
 0.47388157, 0.        , 0.03455471, 0.15020974, 0.40572438],
[0.64292305, 0.75441703, 0.1031392 , 0.54338972, 0.02718516,
 0.54460678, 0.03455471, 0.        , 0.3281208 , 0.63931803],
[0.0687736 , 0.12293897, 0.28066428, 0.25328705, 0.28541859,
 0.49583735, 0.15020974, 0.3281208 , 0.        , 0.11610642],
[0.20058553, 0.01319586, 0.37412884, 0.08426824, 0.48542089,
 0.30505741, 0.40572438, 0.63931803, 0.11610642, 0.        ]])
```

```
dist_sq.shape
```

```
(10, 10)
```

```
dist_sq.diagonal()
```

```
array([0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])
```

3.2.6 Conclusions moving forward

It's important to understand numpy and arrays, since most data sets we encounter are rectangular. The notations and operations we saw in numpy will translate to data, except for the fact that data is typically heterogeneous, i.e., of different types. The problem with using numpy for modern data analysis is that if you have mixed data types, it will all be coerced to strings, and then you can't actually do any data analysis.

The solution to this issue (which is also present in Matlab) came about with the `pandas` package, which is the main workhorse of data science in Python

```
reticulate::use_condaenv('ds', required=T)
```

4 Pandas

4.1 Introduction

`pandas` is the Python Data Analysis package. It allows for data ingestion, transformation and cleaning, and creates objects that can then be passed on to analytic packages like `statsmodels` and `scikit-learn` for modeling and packages like `matplotlib`, `seaborn`, and `plotly` for visualization.

`pandas` is built on top of numpy, so many numpy functions are commonly used in manipulating `pandas` objects.

> `pandas` is a pretty extensive package, and we'll only be able to cover some of its features. For more details, there is free online documentation at pandas.pydata.org. You can also look at the book "Python for Data Analysis (2nd edition)" by Wes McKinney, the original developer of the pandas package, for more details.

4.2 Starting pandas

As with any Python module, you have to "activate" `pandas` by using `import`. The "standard" alias for `pandas` is `pd`. We will also import `numpy`, since `pandas` uses some `numpy` functions in the workflows.

```
import numpy as np
import pandas as pd
```

4.3 Data import and export

Most data sets you will work with are set up in tables, so are rectangular in shape. Think Excel spreadsheets. In `pandas` the structure that will hold this kind of data is a `DataFrame`. We can read external data into a `DataFrame` using one of many `read_*` functions. We can also write from a `DataFrame` to a variety of formats using `to_*` functions. The most common of these are listed below:

Format type	Description	reader	writer
text	CSV	read_csv	to_csv
	Excel	read_excel	to_excel
text	JSON	read_json	to_json
binary	Feather	read_feather	to_feather
binary	SAS	read_sas	
SQL	SQL	read_sql	to_sql

We'll start by reading in the `mtcars` dataset stored as a CSV file

```
pd.read_csv('data/mtcars.csv')
```

```
                 make   mpg  cyl   disp   hp  ...   qsec  vs  am  gear  carb
0           Mazda RX4  21.0    6  160.0  110  ...  16.46   0   1     4     4
1       Mazda RX4 Wag  21.0    6  160.0  110  ...  17.02   0   1     4     4
2          Datsun 710  22.8    4  108.0   93  ...  18.61   1   1     4     1
3      Hornet 4 Drive  21.4    6  258.0  110  ...  19.44   1   0     3     1
4   Hornet Sportabout  18.7    8  360.0  175  ...  17.02   0   0     3     2
5             Valiant  18.1    6  225.0  105  ...  20.22   1   0     3     1
6          Duster 360  14.3    8  360.0  245  ...  15.84   0   0     3     4
7           Merc 240D  24.4    4  146.7   62  ...  20.00   1   0     4     2
8            Merc 230  22.8    4  140.8   95  ...  22.90   1   0     4     2
9            Merc 280  19.2    6  167.6  123  ...  18.30   1   0     4     4
10           Merc 280C  17.8    6  167.6  123  ...  18.90   1   0     4     4
11          Merc 450SE  16.4    8  275.8  180  ...  17.40   0   0     3     3
12          Merc 450SL  17.3    8  275.8  180  ...  17.60   0   0     3     3
13         Merc 450SLC  15.2    8  275.8  180  ...  18.00   0   0     3     3
14  Cadillac Fleetwood  10.4    8  472.0  205  ...  17.98   0   0     3     4
15 Lincoln Continental  10.4    8  460.0  215  ...  17.82   0   0     3     4
16   Chrysler Imperial  14.7    8  440.0  230  ...  17.42   0   0     3     4
17            Fiat 128  32.4    4   78.7   66  ...  19.47   1   1     4     1
18        Honda Civic  30.4    4   75.7   52  ...  18.52   1   1     4     2
19      Toyota Corolla  33.9    4   71.1   65  ...  19.90   1   1     4     1
20       Toyota Corona  21.5    4  120.1   97  ...  20.01   1   0     3     1
21    Dodge Challenger  15.5    8  318.0  150  ...  16.87   0   0     3     2
22         AMC Javelin  15.2    8  304.0  150  ...  17.30   0   0     3     2
23          Camaro Z28  13.3    8  350.0  245  ...  15.41   0   0     3     4
24    Pontiac Firebird  19.2    8  400.0  175  ...  17.05   0   0     3     2
25           Fiat X1-9  27.3    4   79.0   66  ...  18.90   1   1     4     1
26       Porsche 914-2  26.0    4  120.3   91  ...  16.70   0   1     5     2
27        Lotus Europa  30.4    4   95.1  113  ...  16.90   1   1     5     2
28      Ford Pantera L  15.8    8  351.0  264  ...  14.50   0   1     5     4
29        Ferrari Dino  19.7    6  145.0  175  ...  15.50   0   1     5     6
30       Maserati Bora  15.0    8  301.0  335  ...  14.60   0   1     5     8
31          Volvo 142E  21.4    4  121.0  109  ...  18.60   1   1     4     2

[32 rows x 12 columns]
```

This just prints out the data, but then it's lost. To use this data, we have to give it a name, so it's stored in Python's memory

```
mtcars = pd.read_csv('data/mtcars.csv')
```

> One of the big differences between a spreadsheet program and a programming language from the data science perspective is that you have to load data into the programming language. It's

not "just there" like Excel. This is a good thing, since it allows the common functionality of the programming language to work across multiple data sets, and also keeps the original data set pristine. Excel users can run into problems and corrupt their data if they are not careful.

If we wanted to write this data set back out into an Excel file, say, we could do

```
mtcars.to_excel('data/mtcars.xlsx')
```

You may get an error if you don't have the `openpyxl` package installed. You can easily install it from the Anaconda prompt using `conda install openpyxl` and following the prompts.

4.4 Exploring a data set

We would like to get some idea about this data set. There are a bunch of functions linked to the `DataFrame` object that help us in this. First we will use `head` to see the first 8 rows of this data set

```
mtcars.head(8)
```

```
               make   mpg  cyl   disp   hp  ...   qsec  vs  am  gear  carb
0         Mazda RX4  21.0    6  160.0  110  ...  16.46   0   1     4     4
1     Mazda RX4 Wag  21.0    6  160.0  110  ...  17.02   0   1     4     4
2        Datsun 710  22.8    4  108.0   93  ...  18.61   1   1     4     1
3    Hornet 4 Drive  21.4    6  258.0  110  ...  19.44   1   0     3     1
4 Hornet Sportabout  18.7    8  360.0  175  ...  17.02   0   0     3     2
5           Valiant  18.1    6  225.0  105  ...  20.22   1   0     3     1
6        Duster 360  14.3    8  360.0  245  ...  15.84   0   0     3     4
7         Merc 240D  24.4    4  146.7   62  ...  20.00   1   0     4     2

[8 rows x 12 columns]
```

This is our first look into this data. We notice a few things. Each column has a name, and each row has an *index*, starting at 0.

If you're interested in the last N rows, there is a corresponding `tail` function

Let's look at the data types of each of the columns

```
mtcars.dtypes
```

```
make     object
mpg     float64
cyl       int64
disp    float64
hp        int64
drat    float64
```

```
wt      float64
qsec    float64
vs        int64
am        int64
gear      int64
carb      int64
dtype: object
```

This tells us that some of the variables, like `mpg` and `disp`, are floating point (decimal) numbers, several are integers, and `make` is an "object". The `dtypes` function borrows from `numpy`, where there isn't really a type for character or categorical variables. So most often, when you see "object" in the output of `dtypes`, you think it's a character or categorical variable.

We can also look at the data structure in a bit more detail.

```
mtcars.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 12 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   make    32 non-null     object 
 1   mpg     32 non-null     float64
 2   cyl     32 non-null     int64  
 3   disp    32 non-null     float64
 4   hp      32 non-null     int64  
 5   drat    32 non-null     float64
 6   wt      32 non-null     float64
 7   qsec    32 non-null     float64
 8   vs      32 non-null     int64  
 9   am      32 non-null     int64  
 10  gear    32 non-null     int64  
 11  carb    32 non-null     int64  
dtypes: float64(5), int64(6), object(1)
memory usage: 3.1+ KB
```

This tells us that this is indeed a `DataFrame`, wth 12 columns, each with 32 valid observations. Each row has an index value ranging from 0 to 11. We also get the approximate size of this object in memory.

You can also quickly find the number of rows and columns of a data set by using `shape`, which is borrowed from numpy.

```
mtcars.shape
```

```
(32, 12)
```

More generally, we can get a summary of each variable using the `describe` function

```
mtcars.describe()
```

```
             mpg        cyl        disp  ...         am       gear     carb
count  32.000000  32.000000   32.000000  ...  32.000000  32.000000  32.0000
mean   20.090625   6.187500  230.721875  ...   0.406250   3.687500   2.8125
std     6.026948   1.785922  123.938694  ...   0.498991   0.737804   1.6152
min    10.400000   4.000000   71.100000  ...   0.000000   3.000000   1.0000
25%    15.425000   4.000000  120.825000  ...   0.000000   3.000000   2.0000
50%    19.200000   6.000000  196.300000  ...   0.000000   4.000000   2.0000
75%    22.800000   8.000000  326.000000  ...   1.000000   4.000000   4.0000
max    33.900000   8.000000  472.000000  ...   1.000000   5.000000   8.0000

[8 rows x 11 columns]
```

These are usually the first steps in exploring the data.

4.5 Data structures and types

pandas has two main data types: `Series` and `DataFrame`. These are analogous to vectors and matrices, in that a `Series` is 1-dimensional while a `DataFrame` is 2-dimensional.

4.5.1 pandas.Series

The `Series` object holds data from a single input variable, and is required, much like numpy arrays, to be homogeneous in type. You can create `Series` objects from lists or numpy arrays quite easily

```
s = pd.Series([1,3,5,np.nan, 9, 13])
s
```

```
0     1.0
1     3.0
2     5.0
3     NaN
4     9.0
5    13.0
dtype: float64
```

```
s2 = pd.Series(np.arange(1,20))
s2
```

```
0      1
1      2
2      3
3      4
4      5
5      6
6      7
7      8
8      9
9     10
10    11
11    12
12    13
13    14
14    15
15    16
16    17
17    18
18    19
dtype: int64
```

You can access elements of a `Series` much like a `dict`

```
s2[4]
```

```
5
```

There is no requirement that the index of a `Series` has to be numeric. It can be any kind of scalar object

```
s3 = pd.Series(np.random.normal(0,1, (5,)), index = ['a','b','c','d','e'])
s3
```

```
a   -1.059290
b    0.823643
c   -1.593386
d   -1.576311
e   -1.243683
dtype: float64
```

```
s3['d']
```

```
-1.576310998032957
```

```
s3['a':'d']
```

```
a   -1.059290
b    0.823643
c   -1.593386
d   -1.576311
dtype: float64
```

Well, slicing worked, but it gave us something different than expected. It gave us both the start **and** end of the slice, which is unlike what we've encountered so far!!

It turns out that in `pandas`, slicing by index actually does this. It is a discrepancy from `numpy` and Python in general that we have to be careful about.

You can extract the actual values into a numpy array

```
s3.to_numpy()
```

```
array([-1.05929   ,  0.82364332, -1.59338613, -1.576311  , -1.24368268])
```

In fact, you'll see that much of `pandas`' structures are build on top of `numpy` arrays. This is a good thing, since you can take advantage of the powerful numpy functions that are built for fast, efficient scientific computing.

Making the point about slicing again,

```
s3.to_numpy()[0:3]
```

```
array([-1.05929   ,  0.82364332, -1.59338613])
```

This is different from index-based slicing done earlier.

4.5.2 pandas.DataFrame

The `DataFrame` object holds a rectangular data set. Each column of a `DataFrame` is a `Series` object. This means that each column of a `DataFrame` must be comprised of data of the same type, but different columns can hold data of different types. This structure is extremely useful in practical data science. The invention of this structure was, in my opinion, transformative in making Python an effective data science tool.

4.5.2.1 Creating a DataFrame

The `DataFrame` can be created by importing data, as we saw in the previous section. It can also be created by a few methods within Python.

First, it can be created from a 2-dimensional numpy array.

```
rng = np.random.RandomState(25)
d1 = pd.DataFrame(rng.normal(0,1, (4,5)))
d1
```

```
          0         1         2         3         4
0  0.228273  1.026890 -0.839585 -0.591182 -0.956888
1 -0.222326 -0.619915  1.837905 -2.053231  0.868583
2 -0.920734 -0.232312  2.152957 -1.334661  0.076380
3 -1.246089  1.202272 -1.049942  1.056610 -0.419678
```

You will notice that it creates default column names, that are merely the column number, starting from 0. We can also create the column names and row index (similar to the `Series` index we saw earlier) directly during creation.

```
d2 = pd.DataFrame(rng.normal(0,1, (4, 5)),
                  columns = ['A','B','C','D','E'],
                  index = ['a','b','c','d'])
d2
```

```
          A         B         C         D         E
a  2.294842 -2.594487  2.822756  0.680889 -1.577693
b -1.976254  0.533340 -0.290870 -0.513520  1.982626
c  0.226001 -1.839905  1.607671  0.388292  0.399732
d  0.405477  0.217002 -0.633439  0.246622 -1.939546
```

> We could also create a `DataFrame` from a list of lists, as long as things line up, just as we showed for numpy arrays. However, to me, other ways, including the `dict` method below, make more sense.

We can change the column names (which can be extracted and replaced with the `columns` attribute) and the index values (using the `index` attribute).

```
d2.columns
```

```
Index(['A', 'B', 'C', 'D', 'E'], dtype='object')
```

```
d2.columns = pd.Index(['V'+str(i) for i in range(1,6)]) # Index creates the right object
d2
```

```
          V1        V2        V3        V4        V5
a  2.294842 -2.594487  2.822756  0.680889 -1.577693
b -1.976254  0.533340 -0.290870 -0.513520  1.982626
c  0.226001 -1.839905  1.607671  0.388292  0.399732
d  0.405477  0.217002 -0.633439  0.246622 -1.939546
```

Exercise: Can you explain what I did in the list comprehension above? The key points are understanding `str` and how I constructed the `range`.

```
d2.index = ['o1','o2','o3','o4']
d2
```

```
          V1        V2        V3        V4        V5
o1  2.294842 -2.594487  2.822756  0.680889 -1.577693
o2 -1.976254  0.533340 -0.290870 -0.513520  1.982626
o3  0.226001 -1.839905  1.607671  0.388292  0.399732
o4  0.405477  0.217002 -0.633439  0.246622 -1.939546
```

You can also extract data from a homogeneous `DataFrame` to a `numpy` array

```
d1.to_numpy()
```

```
array([[ 0.22827309,  1.0268903 , -0.83958485, -0.59118152, -0.9568883 ],
       [-0.22232569, -0.61991511,  1.83790458, -2.05323076,  0.86858305],
       [-0.92073444, -0.23231186,  2.1529569 , -1.33466147,  0.07637965],
       [-1.24608928,  1.20227231, -1.04994158,  1.05661011, -0.41967767]])
```

> It turns out that you can use `to_numpy` for a non-homogeneous `DataFrame` as well. `numpy` just makes it homogeneous by assigning each column the data type `object`. This also limits what you can do in `numpy` with the array and may require changing data types using the `astype` function. There is some more detail about the `object` data type in the Python Tools for Data Science (notebook, PDF) document.

The other easy way to create a `DataFrame` is from a `dict` object, where each component object is either a list or a numpy array, and is homogeneous in type. One exception is if a component is of size 1; then it is repeated to meet the needs of the `DataFrame`'s dimensions

```
df = pd.DataFrame({
    'A':3.,
    'B':rng.random_sample(5),
    'C': pd.Timestamp('20200512'),
    'D': np.array([6] * 5),
    'E': pd.Categorical(['yes','no','no','yes','no']),
    'F': 'NIH'})
df
```

```
     A         B          C  D    E    F
0  3.0  0.958092 2020-05-12  6  yes  NIH
1  3.0  0.883201 2020-05-12  6   no  NIH
2  3.0  0.295432 2020-05-12  6   no  NIH
3  3.0  0.512376 2020-05-12  6  yes  NIH
4  3.0  0.088702 2020-05-12  6   no  NIH
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   A       5 non-null      float64
 1   B       5 non-null      float64
 2   C       5 non-null      datetime64[ns]
 3   D       5 non-null      int64
 4   E       5 non-null      category
 5   F       5 non-null      object
dtypes: category(1), datetime64[ns](1), float64(2), int64(1), object(1)
memory usage: 429.0+ bytes
```

We note that C is a date object, E is a category object, and F is a text/string object. pandas has excellent time series capabilities (having origins in FinTech), and the `TimeStamp` function creates datetime objects which can be queried and manipulated in Python. We'll describe category data in the next section.

You can also create a `DataFrame` where each column is composed of composite objects, like lists and dicts, as well. This might have limited value in some settings, but may be useful in others. In particular, this allows capabilities like the *list-column* construct in R tibbles. For example,

```
pd.DataFrame({'list' :[[1,2],[3,4],[5,6]],
              'tuple' : [('a','b'), ('c','d'), ('e','f')],
               'set' : [{'A','B','C'}, {'D','E'}, {'F'}],
             'dicts' : [{'A': [1,2,3]}, {'B':[5,6,8]}, {'C': [3,9]}]})
```

```
     list   tuple        set             dicts
0  [1, 2]  (a, b)  {B, C, A}  {'A': [1, 2, 3]}
1  [3, 4]  (c, d)     {E, D}  {'B': [5, 6, 8]}
2  [5, 6]  (e, f)        {F}     {'C': [3, 9]}
```

4.5.2.2 Working with a DataFrame

You can extract particular columns of a `DataFrame` by name

```
df['E']
```

```
0    yes
1     no
2     no
3    yes
4     no
Name: E, dtype: category
Categories (2, object): [no, yes]
```

```
df['B']
```

```
0    0.958092
1    0.883201
2    0.295432
3    0.512376
4    0.088702
Name: B, dtype: float64
```

There is also a shortcut for accessing single columns, using Python's dot (`.`) notation.

```
df.B
```

```
0    0.958092
1    0.883201
2    0.295432
3    0.512376
4    0.088702
Name: B, dtype: float64
```

This notation can be more convenient if we need to perform operations on a single column. If we want to extract multiple columns, this notation will not work. Also, if we want to create new columns or replace existing columns, we need to use the array notation with the column name in quotes.

Let's look at slicing a `DataFrame`

4.5.2.3 Extracting rows and columns

There are two extractor functions in `pandas`:

- `loc` extracts by label (index label, column label, slice of labels, etc.
- `iloc` extracts by index (integers, slice objects, etc.

```
df2 = pd.DataFrame(rng.randint(0,10, (5,4)),
                   index = ['a','b','c','d','e'],
                   columns = ['one','two','three','four'])
df2
```

```
   one  two  three  four
a    5    3      2     8
b    9    3      0     5
c    8    4      3     3
d    5    2      7     1
e    6    7      8     7
```

First, let's see what naively slicing this `DataFrame` does.

```
df2['one']
```

```
a    5
b    9
c    8
d    5
e    6
Name: one, dtype: int64
```

Ok, that works. It grabs one column from the dataset. How about the dot notation?

```
df2.one
```

```
a    5
b    9
c    8
d    5
e    6
Name: one, dtype: int64
```

Let's see what this produces.

```
type(df2.one)
```

```
<class 'pandas.core.series.Series'>
```

So this is a series, so we can potentially do slicing of this series.

```
df2.one['b']
```

```
9
```

```
df2.one['b':'d']
```

```
b    9
c    8
d    5
Name: one, dtype: int64
```

```
df2.one[:3]
```

```
a    5
b    9
c    8
Name: one, dtype: int64
```

Ok, so we have all the `Series` slicing available. The problem here is in semantics, in that we are grabbing one column and then slicing the rows. That doesn't quite work with our sense that a `DataFrame` is a rectangle with rows and columns, and we tend to think of rows, then columns.

Let's see if we can do column slicing with this.

```
df2[:'two']
```

```
   one  two  three  four
a    5    3      2     8
b    9    3      0     5
c    8    4      3     3
d    5    2      7     1
e    6    7      8     7
```

That's not what we want, of course. It's giving back the entire data frame. We'll come back to this.

```
df2[['one','three']]
```

```
   one  three
a    5      2
b    9      0
c    8      3
d    5      7
e    6      8
```

That works correctly though. We can give a list of column names. Ok.

How about row slices?

```
#df2['a'] # Doesn't work
df2['a':'c']
```

```
   one  two  three  four
a    5    3      2     8
b    9    3      0     5
c    8    4      3     3
```

Ok, that works. It slices rows, but includes the largest index, like a `Series` but unlike `numpy` arrays.

```
df2[0:2]
```

```
   one  two  three  four
a    5    3      2     8
b    9    3      0     5
```

Slices by location work too, but use the `numpy` slicing rules.

This entire extraction method becomes confusing. Let's simplify things for this, and then move on to more consistent ways to extract elements of a `DataFrame`. Let's agree on two things. If we're going the direct extraction route,

1. We will extract single columns of a `DataFrame` with `[]` or `.`, i.e., `df2['one']` or `df2.one`
2. We will extract slices of rows of a `DataFrame` using location only, i.e., `df2[:3]`.

For everything else, we'll use two functions, `loc` and `iloc`.

- `loc` extracts elements like a matrix, using index and columns
- `iloc` extracts elements like a matrix, using location

```
df2.loc[:,'one':'three']
```

```
   one  two  three
a    5    3      2
b    9    3      0
c    8    4      3
d    5    2      7
e    6    7      8
```

```
df2.loc['a':'d',:]
```

```
   one  two  three  four
a    5    3      2     8
b    9    3      0     5
c    8    4      3     3
d    5    2      7     1
```

```
df2.loc['b', 'three']
```

```
0
```

So `loc` works just like a matrix, but with `pandas` slicing rules (include largest index)

```
df2.iloc[:,1:4]
```

```
   two  three  four
a    3      2     8
b    3      0     5
c    4      3     3
d    2      7     1
e    7      8     7
```

```
df2.iloc[1:3,:]
```

```
   one  two  three  four
b    9    3      0     5
c    8    4      3     3
```

```
df2.iloc[1:3, 1:4]
```

```
   two  three  four
b    3      0     5
c    4      3     3
```

`iloc` slices like a matrix, but uses `numpy` slicing conventions (does **not** include highest index)

If we want to extract a single element from a dataset, there are two functions available, `iat` and `at`, with behavior corresponding to `iloc` and `loc`, respectively.

```
df2.iat[2,3]
```

```
3
```

```
df2.at['b','three']
```

```
0
```

4.5.2.4 Boolean selection

We can also use tests to extract data from a `DataFrame`. For example, we can extract only rows where column labeled one is greater than 3.

```
df2[df2.one > 3]
```

```
   one  two  three  four
a    5    3      2     8
b    9    3      0     5
c    8    4      3     3
d    5    2      7     1
e    6    7      8     7
```

We can also do composite tests. Here we ask for rows where one is greater than 3 and `three` is less than 9

```
df2[(df2.one > 3) & (df2.three < 9)]
```

```
   one  two  three  four
a    5    3      2     8
b    9    3      0     5
c    8    4      3     3
d    5    2      7     1
e    6    7      8     7
```

4.5.2.5 query

`DataFrame`'s have a `query` method allowing selection using a Python expression

```
n = 10
df = pd.DataFrame(np.random.rand(n, 3), columns = list('abc'))
df
```

```
          a         b         c
0  0.111275  0.483533  0.610500
1  0.435982  0.743773  0.740071
2  0.041377  0.454636  0.455180
3  0.885472  0.131452  0.694800
```

```
4  0.635799  0.047548  0.330058
5  0.666458  0.120961  0.583365
6  0.306277  0.787746  0.394860
7  0.345539  0.379726  0.703836
8  0.832179  0.770601  0.819252
9  0.972577  0.989223  0.486713
```

```
df[(df.a < df.b) & (df.b < df.c)]
```

```
          a         b         c
0  0.111275  0.483533  0.610500
2  0.041377  0.454636  0.455180
7  0.345539  0.379726  0.703836
```

We can equivalently write this query as

```
df.query('(a < b) & (b < c)')
```

```
          a         b         c
0  0.111275  0.483533  0.610500
2  0.041377  0.454636  0.455180
7  0.345539  0.379726  0.703836
```

4.5.2.6 Replacing values in a DataFrame

We can replace values within a DataFrame either by position or using a query.

```
df2
```

```
   one  two  three  four
a    5    3      2     8
b    9    3      0     5
c    8    4      3     3
d    5    2      7     1
e    6    7      8     7
```

```
df2['one'] = [2,5,2,5,2]
df2
```

```
   one  two  three  four
a    2    3      2     8
b    5    3      0     5
c    2    4      3     3
d    5    2      7     1
e    2    7      8     7
```

```
df2.iat[2,3] = -9 # missing value
df2
```

```
   one  two  three  four
a    2    3      2     8
b    5    3      0     5
c    2    4      3    -9
d    5    2      7     1
e    2    7      8     7
```

Let's now replace values using `replace` which is more flexible.

```
df2.replace(0, -9) # replace 0 with -9
```

```
   one  two  three  four
a    2    3      2     8
b    5    3     -9     5
c    2    4      3    -9
d    5    2      7     1
e    2    7      8     7
```

```
df2.replace({2: 2.5, 8: 6.5}) # multiple replacements
```

```
   one  two  three  four
a  2.5  3.0    2.5   6.5
b  5.0  3.0    0.0   5.0
c  2.5  4.0    3.0  -9.0
d  5.0  2.5    7.0   1.0
e  2.5  7.0    6.5   7.0
```

```
df2.replace({'one': {5: 500}, 'three': {0: -9, 8: 800}})
# different replacements in different columns
```

```
   one  two  three  four
a    2    3      2     8
b  500    3     -9     5
c    2    4      3    -9
d  500    2      7     1
e    2    7    800     7
```

See more examples in the documentation

4.5.3 Categorical data

pandas provides a `Categorical` function and a `category` object type to Python. This type is analogous to the `factor` data type in R. It is meant to address categorical or discrete variables, where we need to use them in analyses. Categorical variables typically take on a small number of unique values, like gender, blood type, country of origin, race, etc.

You can create categorical `Series` in a couple of ways:

```
s = pd.Series(['a','b','c'], dtype='category')
```

```
df = pd.DataFrame({
    'A':3.,
    'B':rng.random_sample(5),
    'C': pd.Timestamp('20200512'),
    'D': np.array([6] * 5),
    'E': pd.Categorical(['yes','no','no','yes','no']),
    'F': 'NIH'})
df['F'].astype('category')
```

```
0    NIH
1    NIH
2    NIH
3    NIH
4    NIH
Name: F, dtype: category
Categories (1, object): [NIH]
```

You can also create `DataFrame`'s where each column is categorical

```
df = pd.DataFrame({'A': list('abcd'), 'B': list('bdca')})
df_cat = df.astype('category')
df_cat.dtypes
```

```
A    category
B    category
dtype: object
```

You can explore categorical data in a variety of ways

```
df_cat['A'].describe()
```

```
count     4
unique    4
top       d
freq      1
Name: A, dtype: object
```

```
df['A'].value_counts()
```

```
d    1
a    1
c    1
b    1
Name: A, dtype: int64
```

One issue with categories is that, if a particular level of a category is not seen before, it can create an error. So you can pre-specify the categories you expect

```
df_cat['B'] = pd.Categorical(list('aabb'), categories = ['a','b','c','d'])
df_cat['B'].value_counts()
```

```
b    2
a    2
d    0
c    0
Name: B, dtype: int64
```

4.5.3.1 Re-organizing categories

In categorical data, there is often the concept of a "first" or "reference" category, and an ordering of categories. This tends to be important in both visualization as well as in regression modeling. Both aspects of a category can be addressed using the `reorder_categories` function.

In our earlier example, we can see that the A variable has 4 categories, with the "first" category being "a".

```
df_cat.A
```

```
0    a
1    b
2    c
3    d
Name: A, dtype: category
Categories (4, object): [a, b, c, d]
```

Suppose we want to change this ordering to the reverse ordering, where "d" is the "first" category, and then it goes in reverse order.

```
df_cat['A'] = df_cat.A.cat.reorder_categories(['d','c','b','a'])
df_cat.A
```

```
0    a
1    b
2    c
3    d
Name: A, dtype: category
Categories (4, object): [d, c, b, a]
```

4.5.4 Missing data

Both numpy and pandas allow for missing values, which are a reality in data science. The missing values are coded as `np.nan`. Let's create some data and force some missing values

```
df = pd.DataFrame(np.random.randn(5, 3), index = ['a','c','e', 'f','g'], columns = ['one
df['four'] = 20 # add a column named "four", which will all be 20
df['five'] = df['one'] > 0
df
```

```
        one       two     three  four   five
a  0.488764  0.275880  0.280787    20   True
c  0.193040  0.866247 -0.252054    20   True
e  1.355842 -0.089382  0.269765    20   True
f  1.829710 -0.062615  1.513960    20   True
g -0.425201 -0.596880 -0.518699    20  False
```

```
df2 = df.reindex(['a','b','c','d','e','f','g'])
df2.style.applymap(lambda x: 'background-color:yellow', subset = pd.IndexSlice[['b','d']
```

```
<pandas.io.formats.style.Styler object at 0x13f70d3a0>
```

The code above is creating new blank rows based on the new index values, some of which are present in the existing data and some of which are missing.

We can create *masks* of the data indicating where missing values reside in a data set.

```
df2.isna()
```

```
     one    two  three   four   five
a  False  False  False  False  False
b   True   True   True   True   True
c  False  False  False  False  False
d   True   True   True   True   True
e  False  False  False  False  False
f  False  False  False  False  False
g  False  False  False  False  False
```

```
df2['one'].notna()
```

```
a     True
b    False
c     True
d    False
e     True
f     True
g     True
Name: one, dtype: bool
```

We can obtain complete data by dropping any row that has any missing value. This is called *complete case analysis*, and you should be very careful using it. It is *only* valid if we belive that the missingness is missing at random, and not related to some characteristic of the data or the data gathering process.

```
df2.dropna(how='any')
```

```
        one       two     three  four   five
a  0.488764  0.275880  0.280787  20.0   True
c  0.193040  0.866247 -0.252054  20.0   True
e  1.355842 -0.089382  0.269765  20.0   True
f  1.829710 -0.062615  1.513960  20.0   True
g -0.425201 -0.596880 -0.518699  20.0  False
```

You can also fill in, or *impute*, missing values. This can be done using a single value..

```
out1 = df2.fillna(value = 5)

out1.style.applymap(lambda x: 'background-color:yellow', subset = pd.IndexSlice[['b','d'
```

```
<pandas.io.formats.style.Styler object at 0x13bd33160>
```

or a computed value like a column mean

```
df3 = df2.copy()
df3 = df3.select_dtypes(exclude=[object])   # remove non-numeric columns
out2 = df3.fillna(df3.mean())  # df3.mean() computes column-wise means

out2.style.applymap(lambda x: 'background-color:yellow', subset = pd.IndexSlice[['b','d'
```

```
<pandas.io.formats.style.Styler object at 0x1462e72e0>
```

You can also impute based on the principle of *last value carried forward* which is common in time series. This means that the missing value is imputed with the previous recorded value.

```
out3 = df2.fillna(method = 'ffill') # Fill forward

out3.style.applymap(lambda x: 'background-color:yellow', subset = pd.IndexSlice[['b','d'
```

```
<pandas.io.formats.style.Styler object at 0x14639e790>
```

```
out4 = df2.fillna(method = 'bfill') # Fill backward

out4.style.applymap(lambda x: 'background-color:yellow', subset = pd.IndexSlice[['b','d'
```

```
<pandas.io.formats.style.Styler object at 0x13aac3580>
```

4.6 Data transformation

4.6.1 Arithmetic operations

If you have a `Series` or `DataFrame` that is all numeric, you can add or multiply single numbers to all the elements together.

```
A = pd.DataFrame(np.random.randn(4,5))
print(A)
```

```
          0         1         2         3         4
0 -0.972744 -1.412823 -1.016425 -2.090778 -0.031625
1  0.357643 -1.408205 -2.356258  0.493404  2.156157
2  0.188383  0.542149  1.530398 -0.318730 -0.163390
3 -0.539319 -2.231242 -0.566746  0.108634 -0.270969
```

```
print(A + 6)
```

```
          0         1         2         3         4
0  5.027256  4.587177  4.983575  3.909222  5.968375
1  6.357643  4.591795  3.643742  6.493404  8.156157
2  6.188383  6.542149  7.530398  5.681270  5.836610
3  5.460681  3.768758  5.433254  6.108634  5.729031
```

```
print(A * -10)
```

```
          0          1          2          3          4
0  9.727445  14.128227  10.164255  20.907783   0.316250
1 -3.576427  14.082049  23.562578  -4.934044 -21.561570
2 -1.883830  -5.421495 -15.303979   3.187303   1.633895
3  5.393193  22.312422   5.667460  -1.086337   2.709686
```

If you have two compatible (same dimension) numeric `DataFrames`, you can add, subtract, multiply and divide elementwise

```
B = pd.DataFrame(np.random.randn(4,5) + 4)
print(A + B)
```

```
          0         1         2         3         4
0  3.073781  1.314120  4.082662  2.791453  3.465302
1  2.610519  3.079575  0.978048  5.214583  6.693914
2  3.906562  4.027799  4.946390  3.103448  4.766438
3  4.474672  2.080434  5.385396  4.371403  2.513973
```

```
print(A * B)
```

```
          0         1         2          3         4
0 -3.936235 -3.852686 -5.182843 -10.207663 -0.110590
1  0.805725 -6.319714 -7.856484   2.329450  9.784116
2  0.700442  1.889743  5.227827  -1.090752 -0.805482
3 -2.704142 -9.620393 -3.373353   0.463080 -0.754632
```

If you have a `Series` with the same number of elements as the number of columns of a `DataFrame`, you can do arithmetic operations, with each element of the `Series` acting upon each column of the `DataFrame`

```
c = pd.Series([1,2,3,4,5])
print(A + c)
```

```
          0         1         2         3         4
0  0.027256  0.587177  1.983575  1.909222  4.968375
1  1.357643  0.591795  0.643742  4.493404  7.156157
2  1.188383  2.542149  4.530398  3.681270  4.836610
3  0.460681 -0.231242  2.433254  4.108634  4.729031
```

```
print(A * c)
```

```
          0         1         2         3          4
0 -0.972744 -2.825645 -3.049276 -8.363113  -0.158125
1  0.357643 -2.816410 -7.068774  1.973618  10.780785
2  0.188383  1.084299  4.591194 -1.274921  -0.816948
3 -0.539319 -4.462484 -1.700238  0.434535  -1.354843
```

This idea can be used to standardize a dataset, i.e. make each column have mean 0 and standard deviation 1.

```
means = A.mean(axis=0)
stds = A.std(axis = 0)

(A - means)/stds
```

```
          0         1         2         3         4
0 -1.172324 -0.242093 -0.256896 -1.435313 -0.391565
1  0.960567 -0.238174 -1.087957  0.827844  1.494650
2  0.689208  1.416853  1.322826  0.116598 -0.505167
3 -0.477452 -0.936585  0.022027  0.490872 -0.597917
```

4.6.2 Concatenation of data sets

Let's create some example data sets

```
df1 = pd.DataFrame({'A': ['a'+str(i) for i in range(4)],
    'B': ['b'+str(i) for i in range(4)],
    'C': ['c'+str(i) for i in range(4)],
    'D': ['d'+str(i) for i in range(4)]})

df2 =  pd.DataFrame({'A': ['a'+str(i) for i in range(4,8)],
    'B': ['b'+str(i) for i in range(4,8)],
    'C': ['c'+str(i) for i in range(4,8)],
    'D': ['d'+str(i) for i in range(4,8)]})
df3 =  pd.DataFrame({'A': ['a'+str(i) for i in range(8,12)],
    'B': ['b'+str(i) for i in range(8,12)],
    'C': ['c'+str(i) for i in range(8,12)],
    'D': ['d'+str(i) for i in range(8,12)]})
```

We can concatenate these `DataFrame` objects by row

```
row_concatenate = pd.concat([df1, df2, df3])
print(row_concatenate)
```

```
     A    B    C    D
0   a0   b0   c0   d0
1   a1   b1   c1   d1
2   a2   b2   c2   d2
3   a3   b3   c3   d3
0   a4   b4   c4   d4
1   a5   b5   c5   d5
2   a6   b6   c6   d6
3   a7   b7   c7   d7
0   a8   b8   c8   d8
1   a9   b9   c9   d9
```

```
2  a10  b10  c10  d10
3  a11  b11  c11  d11
```

This stacks the dataframes together. They are literally stacked, as is evidenced by the index values being repeated.

This same exercise can be done by the append function

```
df1.append(df2).append(df3)
```

```
     A    B    C    D
0   a0   b0   c0   d0
1   a1   b1   c1   d1
2   a2   b2   c2   d2
3   a3   b3   c3   d3
0   a4   b4   c4   d4
1   a5   b5   c5   d5
2   a6   b6   c6   d6
3   a7   b7   c7   d7
0   a8   b8   c8   d8
1   a9   b9   c9   d9
2  a10  b10  c10  d10
3  a11  b11  c11  d11
```

Suppose we want to append a new row to `df1`. Lets create a new row.

```
new_row = pd.Series(['n1','n2','n3','n4'])
pd.concat([df1, new_row])
```

```
     A    B    C    D    0
0   a0   b0   c0   d0  NaN
1   a1   b1   c1   d1  NaN
2   a2   b2   c2   d2  NaN
3   a3   b3   c3   d3  NaN
0  NaN  NaN  NaN  NaN   n1
1  NaN  NaN  NaN  NaN   n2
2  NaN  NaN  NaN  NaN   n3
3  NaN  NaN  NaN  NaN   n4
```

That's a lot of missing values. The issue is that the we don't have column names in the `new_row`, and the indices are the same, so pandas tries to append it my making a new column. The solution is to make it a `DataFrame`.

```
new_row = pd.DataFrame([['n1','n2','n3','n4']], columns = ['A','B','C','D'])
print(new_row)
```

```
    A   B   C   D
0  n1  n2  n3  n4
```

```
pd.concat([df1, new_row])
```

```
    A   B   C   D
0  a0  b0  c0  d0
1  a1  b1  c1  d1
2  a2  b2  c2  d2
3  a3  b3  c3  d3
0  n1  n2  n3  n4
```

or

```
df1.append(new_row)
```

```
    A   B   C   D
0  a0  b0  c0  d0
1  a1  b1  c1  d1
2  a2  b2  c2  d2
3  a3  b3  c3  d3
0  n1  n2  n3  n4
```

4.6.2.1 Adding columns

```
pd.concat([df1,df2,df3], axis = 1)
```

```
    A   B   C   D   A   B   C   D    A    B    C    D
0  a0  b0  c0  d0  a4  b4  c4  d4   a8   b8   c8   d8
1  a1  b1  c1  d1  a5  b5  c5  d5   a9   b9   c9   d9
2  a2  b2  c2  d2  a6  b6  c6  d6  a10  b10  c10  d10
3  a3  b3  c3  d3  a7  b7  c7  d7  a11  b11  c11  d11
```

The option `axis=1` ensures that concatenation happens by columns. The default value `axis = 0` concatenates by rows.

Let's play a little game. Let's change the column names of `df2` and `df3` so they are not the same as `df1`.

```
df2.columns = ['E','F','G','H']
df3.columns = ['A','D','F','H']
pd.concat([df1,df2,df3])
```

```
     A    B    C    D    E    F    G    H
0   a0   b0   c0   d0  NaN  NaN  NaN  NaN
1   a1   b1   c1   d1  NaN  NaN  NaN  NaN
2   a2   b2   c2   d2  NaN  NaN  NaN  NaN
3   a3   b3   c3   d3  NaN  NaN  NaN  NaN
0  NaN  NaN  NaN  NaN   a4   b4   c4   d4
1  NaN  NaN  NaN  NaN   a5   b5   c5   d5
2  NaN  NaN  NaN  NaN   a6   b6   c6   d6
3  NaN  NaN  NaN  NaN   a7   b7   c7   d7
0   a8  NaN  NaN   b8  NaN   c8  NaN   d8
1   a9  NaN  NaN   b9  NaN   c9  NaN   d9
2  a10  NaN  NaN  b10  NaN  c10  NaN  d10
3  a11  NaN  NaN  b11  NaN  c11  NaN  d11
```

Now pandas ensures that all column names are represented in the new data frame, but with missing values where the row indices and column indices are mismatched. Some of this can be avoided by only joining on common columns. This is done using the `join` option ir `concat`. The default value is 'outer', which is what you see. above

```
pd.concat([df1, df3], join = 'inner')
```

```
     A    D
0   a0   d0
1   a1   d1
2   a2   d2
3   a3   d3
0   a8   b8
1   a9   b9
2  a10  b10
3  a11  b11
```

You can do the same thing when joining by rows, using `axis = 0` and `join="inner"` to only join on rows with matching indices. Reminder that the indices are just labels and happen to be the row numbers by default.

4.6.3 Merging data sets

For this section we'll use a set of data from a survey, also used by Daniel Chen in "Pandas for Everyone"

```
person = pd.read_csv('data/survey_person.csv')
site = pd.read_csv('data/survey_site.csv')
survey = pd.read_csv('data/survey_survey.csv')
visited = pd.read_csv('data/survey_visited.csv')
```

```
print(person)
```

```
      ident   personal    family
0      dyer    William      Dyer
1        pb      Frank   Pabodie
2      lake   Anderson      Lake
3       roe  Valentina   Roerich
4  danforth      Frank  Danforth
```

```
print(site)
```

```
    name    lat    long
0   DR-1 -49.85 -128.57
1   DR-3 -47.15 -126.72
2  MSK-4 -48.87 -123.40
```

```
print(survey)
```

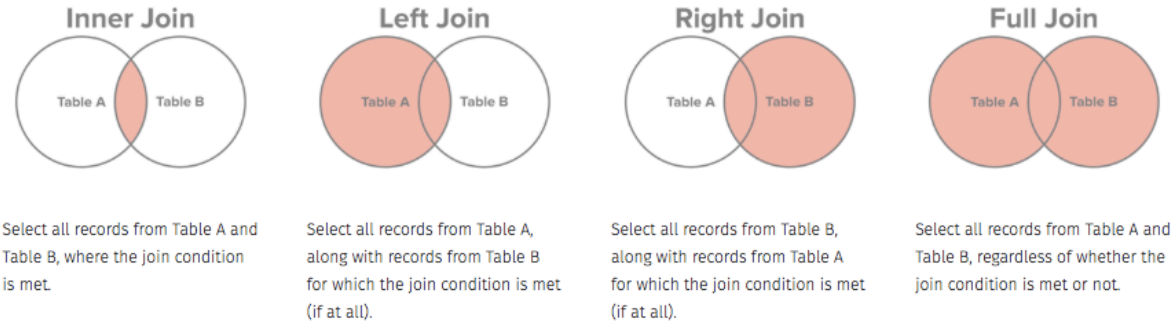
```
    taken person quant  reading
0     619   dyer   rad     9.82
1     619   dyer   sal     0.13
2     622   dyer   rad     7.80
3     622   dyer   sal     0.09
4     734     pb   rad     8.41
5     734   lake   sal     0.05
6     734     pb  temp   -21.50
7     735     pb   rad     7.22
8     735    NaN   sal     0.06
9     735    NaN  temp   -26.00
10    751     pb   rad     4.35
11    751     pb  temp   -18.50
12    751   lake   sal     0.10
13    752   lake   rad     2.19
14    752   lake   sal     0.09
15    752   lake  temp   -16.00
16    752    roe   sal    41.60
17    837   lake   rad     1.46
18    837   lake   sal     0.21
19    837    roe   sal    22.50
20    844    roe   rad    11.25
```

```
print(visited)
```

```
   ident   site       dated
0    619   DR-1  1927-02-08
1    622   DR-1  1927-02-10
2    734   DR-3  1939-01-07
3    735   DR-3  1930-01-12
4    751   DR-3  1930-02-26
5    752   DR-3         NaN
6    837  MSK-4  1932-01-14
7    844   DR-1  1932-03-22
```

There are basically four kinds of joins:

pandas	R	SQL	Description
left	left_join	left outer	keep all rows on left
right	right_join	right outer	keep all rows on right
outer	outer_join	full outer	keep all rows from both
inner	inner_join	inner	keep only rows with common keys



The terms `left` and `right` refer to which data set you call first and second respectively.

We start with an left join

```
s2v_merge = survey.merge(visited, left_on = 'taken',right_on = 'ident', how = 'left')
```

```
print(s2v_merge)
```

```
    taken person quant  reading  ident   site       dated
0     619   dyer   rad     9.82    619   DR-1  1927-02-08
1     619   dyer   sal     0.13    619   DR-1  1927-02-08
2     622   dyer   rad     7.80    622   DR-1  1927-02-10
3     622   dyer   sal     0.09    622   DR-1  1927-02-10
4     734     pb   rad     8.41    734   DR-3  1939-01-07
5     734   lake   sal     0.05    734   DR-3  1939-01-07
6     734     pb  temp   -21.50    734   DR-3  1939-01-07
7     735     pb   rad     7.22    735   DR-3  1930-01-12
```

```
8     735    NaN   sal     0.06    735   DR-3  1930-01-12
9     735    NaN  temp   -26.00    735   DR-3  1930-01-12
10    751     pb   rad     4.35    751   DR-3  1930-02-26
11    751     pb  temp   -18.50    751   DR-3  1930-02-26
12    751   lake   sal     0.10    751   DR-3  1930-02-26
13    752   lake   rad     2.19    752   DR-3         NaN
14    752   lake   sal     0.09    752   DR-3         NaN
15    752   lake  temp   -16.00    752   DR-3         NaN
16    752    roe   sal    41.60    752   DR-3         NaN
17    837   lake   rad     1.46    837  MSK-4  1932-01-14
18    837   lake   sal     0.21    837  MSK-4  1932-01-14
19    837    roe   sal    22.50    837  MSK-4  1932-01-14
20    844    roe   rad    11.25    844   DR-1  1932-03-22
```

Here, the left dataset is `survey` and the right one is `visited`. Since we're doing a left join, we keed all the rows from `survey` and add columns from `visited`, matching on the common key, called "taken" in one dataset and "ident" in the other. Note that the rows of `visited` are repeated as needed to line up with all the rows with common "taken" values.

We can now add location information, where the common key is the site code

```
s2v2loc_merge = s2v_merge.merge(site, how = 'left', left_on = 'site', right_on = 'name')
print(s2v2loc_merge)
```

```
    taken person quant  reading  ident   site       dated   name    lat    long
0     619   dyer   rad     9.82    619   DR-1  1927-02-08   DR-1 -49.85 -128.57
1     619   dyer   sal     0.13    619   DR-1  1927-02-08   DR-1 -49.85 -128.57
2     622   dyer   rad     7.80    622   DR-1  1927-02-10   DR-1 -49.85 -128.57
3     622   dyer   sal     0.09    622   DR-1  1927-02-10   DR-1 -49.85 -128.57
4     734     pb   rad     8.41    734   DR-3  1939-01-07   DR-3 -47.15 -126.72
5     734   lake   sal     0.05    734   DR-3  1939-01-07   DR-3 -47.15 -126.72
6     734     pb  temp   -21.50    734   DR-3  1939-01-07   DR-3 -47.15 -126.72
7     735     pb   rad     7.22    735   DR-3  1930-01-12   DR-3 -47.15 -126.72
8     735    NaN   sal     0.06    735   DR-3  1930-01-12   DR-3 -47.15 -126.72
9     735    NaN  temp   -26.00    735   DR-3  1930-01-12   DR-3 -47.15 -126.72
10    751     pb   rad     4.35    751   DR-3  1930-02-26   DR-3 -47.15 -126.72
11    751     pb  temp   -18.50    751   DR-3  1930-02-26   DR-3 -47.15 -126.72
12    751   lake   sal     0.10    751   DR-3  1930-02-26   DR-3 -47.15 -126.72
13    752   lake   rad     2.19    752   DR-3         NaN   DR-3 -47.15 -126.72
14    752   lake   sal     0.09    752   DR-3         NaN   DR-3 -47.15 -126.72
15    752   lake  temp   -16.00    752   DR-3         NaN   DR-3 -47.15 -126.72
16    752    roe   sal    41.60    752   DR-3         NaN   DR-3 -47.15 -126.72
17    837   lake   rad     1.46    837  MSK-4  1932-01-14  MSK-4 -48.87 -123.40
18    837   lake   sal     0.21    837  MSK-4  1932-01-14  MSK-4 -48.87 -123.40
19    837    roe   sal    22.50    837  MSK-4  1932-01-14  MSK-4 -48.87 -123.40
20    844    roe   rad    11.25    844   DR-1  1932-03-22   DR-1 -49.85 -128.57
```

Lastly, we add the person information to this dataset.

```
merged = s2v2loc_merge.merge(person, how = 'left', left_on = 'person', right_on = 'ident
print(merged.head())
```

```
   taken person quant  reading  ...    long ident_y personal   family
0    619   dyer   rad     9.82  ... -128.57    dyer  William     Dyer
1    619   dyer   sal     0.13  ... -128.57    dyer  William     Dyer
2    622   dyer   rad     7.80  ... -128.57    dyer  William     Dyer
3    622   dyer   sal     0.09  ... -128.57    dyer  William     Dyer
4    734     pb   rad     8.41  ... -126.72      pb    Frank  Pabodie

[5 rows x 13 columns]
```

You can merge based on multiple columns as long as they match up.

```
ps = person.merge(survey, left_on = 'ident', right_on = 'person')
vs = visited.merge(survey, left_on = 'ident', right_on = 'taken')
print(ps)
```

```
   ident   personal   family  taken person quant  reading
0   dyer    William     Dyer    619   dyer   rad     9.82
1   dyer    William     Dyer    619   dyer   sal     0.13
2   dyer    William     Dyer    622   dyer   rad     7.80
3   dyer    William     Dyer    622   dyer   sal     0.09
4     pb      Frank  Pabodie    734     pb   rad     8.41
5     pb      Frank  Pabodie    734     pb  temp   -21.50
6     pb      Frank  Pabodie    735     pb   rad     7.22
7     pb      Frank  Pabodie    751     pb   rad     4.35
8     pb      Frank  Pabodie    751     pb  temp   -18.50
9   lake   Anderson     Lake    734   lake   sal     0.05
10  lake   Anderson     Lake    751   lake   sal     0.10
11  lake   Anderson     Lake    752   lake   rad     2.19
12  lake   Anderson     Lake    752   lake   sal     0.09
13  lake   Anderson     Lake    752   lake  temp   -16.00
14  lake   Anderson     Lake    837   lake   rad     1.46
15  lake   Anderson     Lake    837   lake   sal     0.21
16   roe  Valentina  Roerich    752    roe   sal    41.60
17   roe  Valentina  Roerich    837    roe   sal    22.50
18   roe  Valentina  Roerich    844    roe   rad    11.25
```

```
print(vs)
```

```
    ident   site       dated  taken person quant  reading
0     619   DR-1  1927-02-08    619   dyer   rad     9.82
```

```
1     619   DR-1  1927-02-08    619   dyer   sal     0.13
2     622   DR-1  1927-02-10    622   dyer   rad     7.80
3     622   DR-1  1927-02-10    622   dyer   sal     0.09
4     734   DR-3  1939-01-07    734     pb   rad     8.41
5     734   DR-3  1939-01-07    734   lake   sal     0.05
6     734   DR-3  1939-01-07    734     pb  temp   -21.50
7     735   DR-3  1930-01-12    735     pb   rad     7.22
8     735   DR-3  1930-01-12    735    NaN   sal     0.06
9     735   DR-3  1930-01-12    735    NaN  temp   -26.00
10    751   DR-3  1930-02-26    751     pb   rad     4.35
11    751   DR-3  1930-02-26    751     pb  temp   -18.50
12    751   DR-3  1930-02-26    751   lake   sal     0.10
13    752   DR-3         NaN    752   lake   rad     2.19
14    752   DR-3         NaN    752   lake   sal     0.09
15    752   DR-3         NaN    752   lake  temp   -16.00
16    752   DR-3         NaN    752    roe   sal    41.60
17    837  MSK-4  1932-01-14    837   lake   rad     1.46
18    837  MSK-4  1932-01-14    837   lake   sal     0.21
19    837  MSK-4  1932-01-14    837    roe   sal    22.50
20    844   DR-1  1932-03-22    844    roe   rad    11.25
```

```
ps_vs = ps.merge(vs,
                 left_on = ['ident','taken', 'quant','reading'],
                 right_on = ['person','ident','quant','reading']) # The keys need to corr
ps_vs.head()
```

```
  ident_x personal   family  taken_x  ...  site       dated  taken_y  person_y
0    dyer  William     Dyer      619  ...  DR-1  1927-02-08      619      dyer
1    dyer  William     Dyer      619  ...  DR-1  1927-02-08      619      dyer
2    dyer  William     Dyer      622  ...  DR-1  1927-02-10      622      dyer
3    dyer  William     Dyer      622  ...  DR-1  1927-02-10      622      dyer
4      pb    Frank  Pabodie      734  ...  DR-3  1939-01-07      734        pb

[5 rows x 12 columns]
```

Note that since there are common column names, the merge appends _x and _y to denote which column came from the left and right, respectively.

4.6.4 Tidy data principles and reshaping datasets

The tidy data principle is a principle espoused by Dr. Hadley Wickham, one of the foremost R developers. Tidy data is a structure for datasets to make them more easily analyzed on computers. The basic principles are

- Each row is an observation
- Each column is a variable

- Each type of observational unit forms a table

  Tidy data is tidy in one way. Untidy data can be untidy in many ways

Let's look at some examples.

```
from glob import glob
filenames = sorted(glob('data/table*.csv')) # find files matching pattern. I know there
table1, table2, table3, table4a, table4b, table5 = [pd.read_csv(f) for f in filenames] #
```

This code imports data from 6 files matching a pattern. Python allows multiple assignments on the left of the =, and as each dataset is imported, it gets assigned in order to the variables on the left. In the second line I sort the file names so that they match the order in which I'm storing them in the 3rd line. The function `glob` does pattern-matching of file names.

The following tables refer to the number of TB cases and population in Afghanistan, Brazil and China in 1999 and 2000

```
print(table1)
```

```
       country  year   cases  population
0  Afghanistan  1999     745    19987071
1  Afghanistan  2000    2666    20595360
2       Brazil  1999   37737   172006362
3       Brazil  2000   80488   174504898
4        China  1999  212258  1272915272
5        China  2000  213766  1280428583
```

```
print(table2)
```

```
        country  year        type       count
0   Afghanistan  1999       cases         745
1   Afghanistan  1999  population    19987071
2   Afghanistan  2000       cases        2666
3   Afghanistan  2000  population    20595360
4        Brazil  1999       cases       37737
5        Brazil  1999  population   172006362
6        Brazil  2000       cases       80488
7        Brazil  2000  population   174504898
8         China  1999       cases      212258
9         China  1999  population  1272915272
10        China  2000       cases      213766
11        China  2000  population  1280428583
```

```
print(table3)
```

```
       country  year               rate
0  Afghanistan  1999       745/19987071
1  Afghanistan  2000      2666/20595360
2       Brazil  1999    37737/172006362
3       Brazil  2000    80488/174504898
4        China  1999  212258/1272915272
5        China  2000  213766/1280428583
```

```
print(table4a) # cases
```

```
       country    1999    2000
0  Afghanistan     745    2666
1       Brazil   37737   80488
2        China  212258  213766
```

```
print(table4b) # population
```

```
       country        1999        2000
0  Afghanistan    19987071    20595360
1       Brazil   172006362   174504898
2        China  1272915272  1280428583
```

```
print(table5)
```

```
       country  century  year               rate
0  Afghanistan       19    99       745/19987071
1  Afghanistan       20     0      2666/20595360
2       Brazil       19    99    37737/172006362
3       Brazil       20     0    80488/174504898
4        China       19    99  212258/1272915272
5        China       20     0  213766/1280428583
```

Exercise: Describe why and why not each of these datasets are tidy.

4.6.5 Melting (unpivoting) data

Melting is the operation of collapsing multiple columns into 2 columns, where one column is formed by the old column names, and the other by the corresponding values. Some columns may be kept fixed and their data are repeated to maintain the interrelationships between the variables.

We'll start with loading some data on income and religion in the US from the Pew Research Center.

```
pew = pd.read_csv('data/pew.csv')
print(pew.head())
```

```
            religion  <$10k  $10-20k  ...  $100-150k  >150k  Don't know/refused
0           Agnostic     27       34  ...        109     84                  96
1            Atheist     12       27  ...         59     74                  76
2           Buddhist     27       21  ...         39     53                  54
3           Catholic    418      617  ...        792    633                1489
4  Don't know/refused     15       14  ...         17     18                 116

[5 rows x 11 columns]
```

This dataset is considered in "wide" format. There are several issues with it, including the fact that column headers have data. Those column headers are income groups, that should be a column by tidy principles. Our job is to turn this dataset into "long" format with a column for income group.

We will use the function `melt` to achieve this. This takes a few parameters:

- **id_vars** is a list of variables that will remain as is
- **value_vars** is a list of column nmaes that we will melt (or unpivot). By default, it will melt all columns not mentioned in id_vars
- **var_name** is a string giving the name of the new column created by the headers (default: `variable`)
- **value_name** is a string giving the name of the new column created by the values (default: `value`)

```
pew_long = pew.melt(id_vars = ['religion'], var_name = 'income_group', value_name = 'cou
print(pew_long.head())
```

```
             religion income_group  count
0            Agnostic        <$10k     27
1             Atheist        <$10k     12
2            Buddhist        <$10k     27
3            Catholic        <$10k    418
4  Don't know/refused        <$10k     15
```

4.6.6 Separating columns containing multiple variables

We will use an Ebola dataset to illustrate this principle

```
ebola = pd.read_csv('data/country_timeseries.csv')
print(ebola.head())
```

```
         Date  Day  ...  Deaths_Spain  Deaths_Mali
0    1/5/2015  289  ...           NaN          NaN
1    1/4/2015  288  ...           NaN          NaN
2    1/3/2015  287  ...           NaN          NaN
3    1/2/2015  286  ...           NaN          NaN
4  12/31/2014  284  ...           NaN          NaN

[5 rows x 18 columns]
```

Note that for each country we have two columns – one for cases (number infected) and one for deaths. Ideally we want one column for country, one for cases and one for deaths.

The first step will be to melt this data sets so that the column headers in question from a column and the corresponding data forms a second column.

```
ebola_long = ebola.melt(id_vars = ['Date','Day'])
print(ebola_long.head())
```

```
         Date  Day      variable   value
0    1/5/2015  289  Cases_Guinea  2776.0
1    1/4/2015  288  Cases_Guinea  2775.0
2    1/3/2015  287  Cases_Guinea  2769.0
3    1/2/2015  286  Cases_Guinea     NaN
4  12/31/2014  284  Cases_Guinea  2730.0
```

We now need to split the data in the `variable` column to make two columns. One will contain the country name and the other either Cases or Deaths. We will use some string manipulation functions that we will see later to achieve this.

```
variable_split = ebola_long['variable'].str.split('_', expand=True) # split on the `_` c
print(variable_split[:5])
```

```
       0       1
0  Cases  Guinea
1  Cases  Guinea
2  Cases  Guinea
3  Cases  Guinea
4  Cases  Guinea
```

The `expand=True` option forces the creation of an `DataFrame` rather than a list

```
type(variable_split)
```

```
<class 'pandas.core.frame.DataFrame'>
```

We can now concatenate this to the original data

```
variable_split.columns = ['status','country']

ebola_parsed = pd.concat([ebola_long, variable_split], axis = 1)

ebola_parsed.drop('variable', axis = 1, inplace=True) # Remove the column named "variabl

print(ebola_parsed.head())
```

```
         Date  Day   value status country
0    1/5/2015  289  2776.0  Cases  Guinea
1    1/4/2015  288  2775.0  Cases  Guinea
2    1/3/2015  287  2769.0  Cases  Guinea
3    1/2/2015  286     NaN  Cases  Guinea
4  12/31/2014  284  2730.0  Cases  Guinea
```

4.6.7 Pivot/spread datasets

If we wanted to, we could also make two columns based on cases and deaths, so for each country and date you could easily read off the cases and deaths. This is achieved using the `pivot_table` function.

In the `pivot_table` syntax, `index` refers to the columns we don't want to change, `columns` refers to the column whose values will form the column names of the new columns, and `values` is the name of the column that will form the values in the pivoted dataset.

```
ebola_parsed.pivot_table(index = ['Date','Day', 'country'], columns = 'status', values =
```

```
status                    Cases  Deaths
Date     Day country
1/2/2015 286 Liberia     8157.0  3496.0
1/3/2015 287 Guinea      2769.0  1767.0
             Liberia     8166.0  3496.0
             SierraLeone 9722.0  2915.0
1/4/2015 288 Guinea      2775.0  1781.0
...                         ...     ...
9/7/2014 169 Liberia     2081.0  1137.0
             Nigeria       21.0     8.0
             Senegal        3.0     0.0
             SierraLeone  1424.0   524.0
9/9/2014 171 Liberia     2407.0     NaN

[375 rows x 2 columns]
```

This creates something called `MultiIndex` in the pandas `DataFrame`. This is useful in some advanced cases, but here, we just want a normal `DataFrame` back. We can achieve that by using the `reset_index` function.

```
ebola_parsed.pivot_table(index = ['Date','Day','country'], columns = 'status', values =
```

```
status      Date  Day      country   Cases  Deaths
0       1/2/2015  286      Liberia  8157.0  3496.0
1       1/3/2015  287       Guinea  2769.0  1767.0
2       1/3/2015  287      Liberia  8166.0  3496.0
3       1/3/2015  287  SierraLeone  9722.0  2915.0
4       1/4/2015  288       Guinea  2775.0  1781.0
```

```
..           ...  ...          ...     ...     ...
370     9/7/2014  169      Liberia  2081.0  1137.0
371     9/7/2014  169      Nigeria    21.0     8.0
372     9/7/2014  169      Senegal     3.0     0.0
373     9/7/2014  169  SierraLeone  1424.0   524.0
374     9/9/2014  171      Liberia  2407.0     NaN

[375 rows x 5 columns]
```

Pivoting is a 2-column to many-column operation, with the number of columns formed depending on the number of unique values present in the column of the original data that is entered into the `columns` argument of `pivot_table`

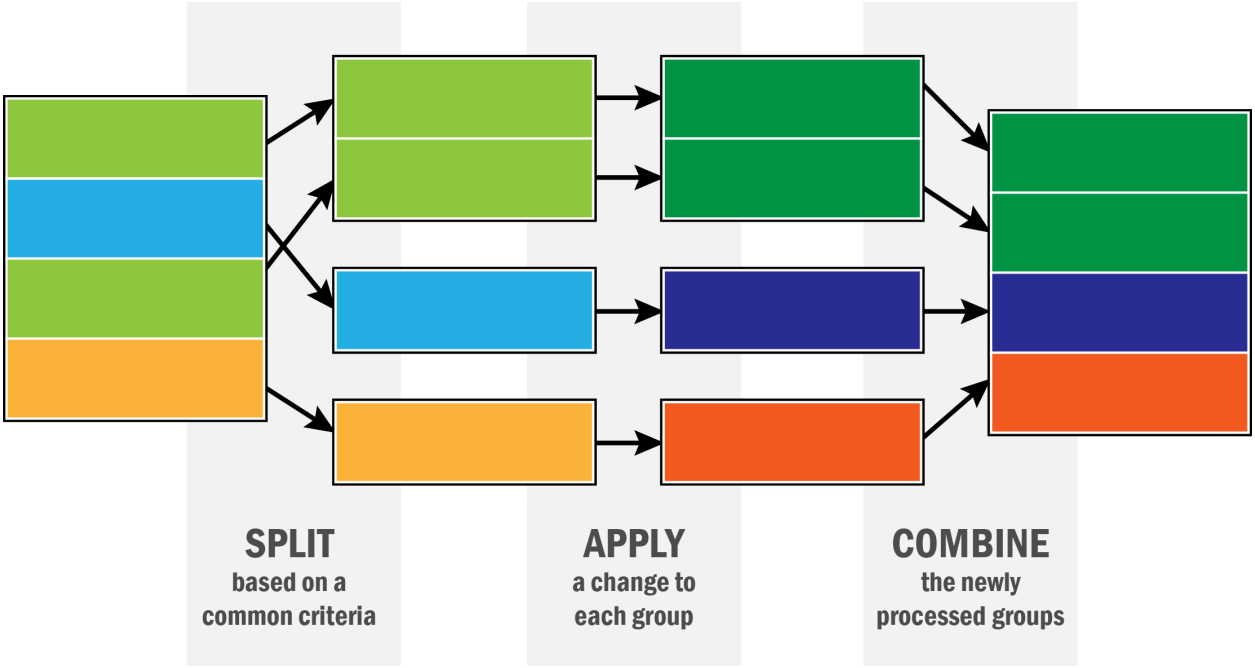
Exercise: Load the file `weather.csv` into Python and work on making it a tidy dataset. It requires melting and pivoting. The dataset comprises of the maximun and minimum temperatures recorded each day in 2010. There are lots of missing value. Ultimately we want columns for days of the month, maximum temperature and minimum tempearture along with the location ID, the year and the month.

4.7 Data aggregation and split-apply-combine

We'll use the Gapminder dataset for this section

```
df = pd.read_csv('data/gapminder.tsv', sep = '\t') # data is tab-separated, so we use `\
```

The paradigm we will be exploring is often called *split-apply-combine* or MapReduce or grouped aggregation. The basic idea is that you split a data set up by some feature, apply a recipe to each piece, compute the result, and then put the results back together into a dataset. This can be described in teh following schematic.



pandas is set up for this. It features the groupby function that allows the "split" part of the operation. We can then apply a function to each part and put it back together. Let's see how.

```
df.head()
```

```
       country continent  year  lifeExp       pop   gdpPercap
0  Afghanistan      Asia  1952   28.801   8425333  779.445314
1  Afghanistan      Asia  1957   30.332   9240934  820.853030
2  Afghanistan      Asia  1962   31.997  10267083  853.100710
3  Afghanistan      Asia  1967   34.020  11537966  836.197138
4  Afghanistan      Asia  1972   36.088  13079460  739.981106
```

```
f"This dataset has {len(df['country'].unique())} countries in it"
```

```
'This dataset has 142 countries in it'
```

One of the variables in this dataset is life expectancy at birth, lifeExp. Suppose we want to find the average life expectancy of each country over the period of study.

```
df.groupby('country')['lifeExp'].mean()
```

```
country
Afghanistan           37.478833
Albania               68.432917
Algeria               59.030167
Angola                37.883500
Argentina             69.060417
                          ...
Vietnam               57.479500
West Bank and Gaza    60.328667
Yemen, Rep.           46.780417
Zambia                45.996333
Zimbabwe              52.663167
Name: lifeExp, Length: 142, dtype: float64
```

So what's going on here? First, we use the groupby function, telling pandas to split the dataset up by values of the column country.

```
df.groupby('country')
```

```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x13f6fc1c0>
```

pandas won't show you the actual data, but will tell you that it is a grouped dataframe object. This means that each element of this object is a DataFrame with data from one country.

```
df.groupby('country').ngroups
```

```
142
```

```
df.groupby('country').get_group('United Kingdom')
```

```
             country continent  year  lifeExp       pop     gdpPercap
1596  United Kingdom    Europe  1952   69.180  50430000   9979.508487
1597  United Kingdom    Europe  1957   70.420  51430000  11283.177950
1598  United Kingdom    Europe  1962   70.760  53292000  12477.177070
1599  United Kingdom    Europe  1967   71.360  54959000  14142.850890
1600  United Kingdom    Europe  1972   72.010  56079000  15895.116410
1601  United Kingdom    Europe  1977   72.760  56179000  17428.748460
1602  United Kingdom    Europe  1982   74.040  56339704  18232.424520
1603  United Kingdom    Europe  1987   75.007  56981620  21664.787670
1604  United Kingdom    Europe  1992   76.420  57866349  22705.092540
1605  United Kingdom    Europe  1997   77.218  58808266  26074.531360
1606  United Kingdom    Europe  2002   78.471  59912431  29478.999190
1607  United Kingdom    Europe  2007   79.425  60776238  33203.261280
```

```
type(df.groupby('country').get_group('United Kingdom'))
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
avg_lifeexp_country = df.groupby('country').lifeExp.mean()
avg_lifeexp_country['United Kingdom']
```

```
73.92258333333332
```

```
df.groupby('country').get_group('United Kingdom').lifeExp.mean()
```

```
73.92258333333332
```

Let's look at if life expectancy has gone up over time, by continent

```
df.groupby(['continent','year']).lifeExp.mean()
```

```
continent  year
Africa     1952    39.135500
           1957    41.266346
           1962    43.319442
           1967    45.334538
```

```
            1972    47.450942
            1977    49.580423
            1982    51.592865
            1987    53.344788
            1992    53.629577
            1997    53.598269
            2002    53.325231
            2007    54.806038
Americas    1952    53.279840
            1957    55.960280
            1962    58.398760
            1967    60.410920
            1972    62.394920
            1977    64.391560
            1982    66.228840
            1987    68.090720
            1992    69.568360
            1997    71.150480
            2002    72.422040
            2007    73.608120
Asia        1952    46.314394
            1957    49.318544
            1962    51.563223
            1967    54.663640
            1972    57.319269
            1977    59.610556
            1982    62.617939
            1987    64.851182
            1992    66.537212
            1997    68.020515
            2002    69.233879
            2007    70.728485
Europe      1952    64.408500
            1957    66.703067
            1962    68.539233
            1967    69.737600
            1972    70.775033
            1977    71.937767
            1982    72.806400
            1987    73.642167
            1992    74.440100
            1997    75.505167
            2002    76.700600
            2007    77.648600
Oceania     1952    69.255000
            1957    70.295000
            1962    71.085000
            1967    71.310000
```

```
            1972    71.910000
            1977    72.855000
            1982    74.290000
            1987    75.320000
            1992    76.945000
            1997    78.190000
            2002    79.740000
            2007    80.719500
Name: lifeExp, dtype: float64
```

```
avg_lifeexp_continent_yr = df.groupby(['continent','year']).lifeExp.mean().reset_index()
avg_lifeexp_continent_yr
```

```
   continent  year    lifeExp
0     Africa  1952  39.135500
1     Africa  1957  41.266346
2     Africa  1962  43.319442
3     Africa  1967  45.334538
4     Africa  1972  47.450942
5     Africa  1977  49.580423
6     Africa  1982  51.592865
7     Africa  1987  53.344788
8     Africa  1992  53.629577
9     Africa  1997  53.598269
10    Africa  2002  53.325231
11    Africa  2007  54.806038
12  Americas  1952  53.279840
13  Americas  1957  55.960280
14  Americas  1962  58.398760
15  Americas  1967  60.410920
16  Americas  1972  62.394920
17  Americas  1977  64.391560
18  Americas  1982  66.228840
19  Americas  1987  68.090720
20  Americas  1992  69.568360
21  Americas  1997  71.150480
22  Americas  2002  72.422040
23  Americas  2007  73.608120
24      Asia  1952  46.314394
25      Asia  1957  49.318544
26      Asia  1962  51.563223
27      Asia  1967  54.663640
28      Asia  1972  57.319269
29      Asia  1977  59.610556
30      Asia  1982  62.617939
31      Asia  1987  64.851182
32      Asia  1992  66.537212
```

```
33      Asia  1997  68.020515
34      Asia  2002  69.233879
35      Asia  2007  70.728485
36    Europe  1952  64.408500
37    Europe  1957  66.703067
38    Europe  1962  68.539233
39    Europe  1967  69.737600
40    Europe  1972  70.775033
41    Europe  1977  71.937767
42    Europe  1982  72.806400
43    Europe  1987  73.642167
44    Europe  1992  74.440100
45    Europe  1997  75.505167
46    Europe  2002  76.700600
47    Europe  2007  77.648600
48   Oceania  1952  69.255000
49   Oceania  1957  70.295000
50   Oceania  1962  71.085000
51   Oceania  1967  71.310000
52   Oceania  1972  71.910000
53   Oceania  1977  72.855000
54   Oceania  1982  74.290000
55   Oceania  1987  75.320000
56   Oceania  1992  76.945000
57   Oceania  1997  78.190000
58   Oceania  2002  79.740000
59   Oceania  2007  80.719500
```

```
type(avg_lifeexp_continent_yr)
```

```
<class 'pandas.core.frame.DataFrame'>
```

The aggregation function, in this case `mean`, does both the "apply" and "combine" parts of the process.

We can do quick aggregations with `pandas`

```
df.groupby('continent').lifeExp.describe()
```

```
           count       mean        std  ...      50%       75%     max
continent                               ...
Africa     624.0  48.865330   9.150210  ...  47.7920  54.41150  76.442
Americas   300.0  64.658737   9.345088  ...  67.0480  71.69950  80.653
Asia       396.0  60.064903  11.864532  ...  61.7915  69.50525  82.603
Europe     360.0  71.903686   5.433178  ...  72.2410  75.45050  81.757
Oceania     24.0  74.326208   3.795611  ...  73.6650  77.55250  81.235

[5 rows x 8 columns]
```

```
df.groupby('continent').nth(10) # Tenth observation in each group
```

```
               country  year  lifeExp       pop     gdpPercap
continent
Africa         Algeria  2002   70.994  31287142   5288.040382
Americas     Argentina  2002   74.340  38331121   8797.640716
Asia       Afghanistan  2002   42.129  25268405    726.734055
Europe         Albania  2002   75.651   3508512   4604.211737
Oceania      Australia  2002   80.370  19546792  30687.754730
```

You can also use functions from other modules, or your own functions in this aggregation work.

```
df.groupby('continent').lifeExp.agg(np.mean)
```

```
continent
Africa      48.865330
Americas    64.658737
Asia        60.064903
Europe      71.903686
Oceania     74.326208
Name: lifeExp, dtype: float64
```

```
def my_mean(values):
    n = len(values)
    sum = 0
    for value in values:
        sum += value
    return(sum/n)

df.groupby('continent').lifeExp.agg(my_mean)
```

```
continent
Africa      48.865330
Americas    64.658737
Asia        60.064903
Europe      71.903686
Oceania     74.326208
Name: lifeExp, dtype: float64
```

You can do many functions at once

```
df.groupby('year').lifeExp.agg([np.count_nonzero, np.mean, np.std])
```

```
      count_nonzero       mean        std
year
1952          142.0  49.057620  12.225956
1957          142.0  51.507401  12.231286
1962          142.0  53.609249  12.097245
1967          142.0  55.678290  11.718858
1972          142.0  57.647386  11.381953
1977          142.0  59.570157  11.227229
1982          142.0  61.533197  10.770618
1987          142.0  63.212613  10.556285
1992          142.0  64.160338  11.227380
1997          142.0  65.014676  11.559439
2002          142.0  65.694923  12.279823
2007          142.0  67.007423  12.073021
```

You can also aggregate on different columns at the same time by passing a `dict` to the `agg` function

```
df.groupby('year').agg({'lifeExp': np.mean,'pop': np.median,'gdpPercap': np.median}).res
```

```
    year    lifeExp         pop    gdpPercap
0   1952  49.057620   3943953.0  1968.528344
1   1957  51.507401   4282942.0  2173.220291
2   1962  53.609249   4686039.5  2335.439533
3   1967  55.678290   5170175.5  2678.334741
4   1972  57.647386   5877996.5  3339.129407
5   1977  59.570157   6404036.5  3798.609244
6   1982  61.533197   7007320.0  4216.228428
7   1987  63.212613   7774861.5  4280.300366
8   1992  64.160338   8688686.5  4386.085502
9   1997  65.014676   9735063.5  4781.825478
10  2002  65.694923  10372918.5  5319.804524
11  2007  67.007423  10517531.0  6124.371109
```

4.7.0.1 Transformation

You can do grouped transformations using this same method. We will compute the z-score for each year, i.e. we will substract the average life expectancy and divide by the standard deviation

```
def my_zscore(values):
    m = np.mean(values)
    s = np.std(values)
    return((values - m)/s)
```

```
df.groupby('year').lifeExp.transform(my_zscore)
```

```
0      -1.662719
1      -1.737377
2      -1.792867
3      -1.854699
4      -1.900878
          ...
1699   -0.081910
1700   -0.338167
1701   -1.580537
1702   -2.100756
1703   -1.955077
Name: lifeExp, Length: 1704, dtype: float64
```

```
df['lifeExp_z'] = df.groupby('year').lifeExp.transform(my_zscore)
```

```
df.groupby('year').lifeExp_z.mean()
```

```
year
1952   -5.165078e-16
1957    2.902608e-17
1962    2.404180e-16
1967   -6.108181e-16
1972    1.784566e-16
1977   -9.456442e-16
1982   -1.623310e-16
1987    6.687725e-16
1992    5.457293e-16
1997    8.787963e-16
2002    5.254013e-16
2007    4.925637e-16
Name: lifeExp_z, dtype: float64
```

4.7.0.2 Filter

We can split the dataset by values of one variable, and filter out those splits that fail some criterion. The following code only keeps countries with a population of at least 10 million at some point during the study period

```
df.groupby('country').filter(lambda d: d['pop'].max() > 10000000)
```

	country	continent	year	lifeExp	pop	gdpPercap	lifeExp_z
0	Afghanistan	Asia	1952	28.801	8425333	779.445314	-1.662719
1	Afghanistan	Asia	1957	30.332	9240934	820.853030	-1.737377
2	Afghanistan	Asia	1962	31.997	10267083	853.100710	-1.792867
3	Afghanistan	Asia	1967	34.020	11537966	836.197138	-1.854699

```
4     Afghanistan      Asia  1972   36.088  13079460  739.981106  -1.900878
...           ...       ...   ...      ...       ...         ...        ...
1699     Zimbabwe    Africa  1987   62.351   9216418  706.157306  -0.081910
1700     Zimbabwe    Africa  1992   60.377  10704340  693.420786  -0.338167
1701     Zimbabwe    Africa  1997   46.809  11404948  792.449960  -1.580537
1702     Zimbabwe    Africa  2002   39.989  11926563  672.038623  -2.100756
1703     Zimbabwe    Africa  2007   43.487  12311143  469.709298  -1.955077

[924 rows x 7 columns]
```

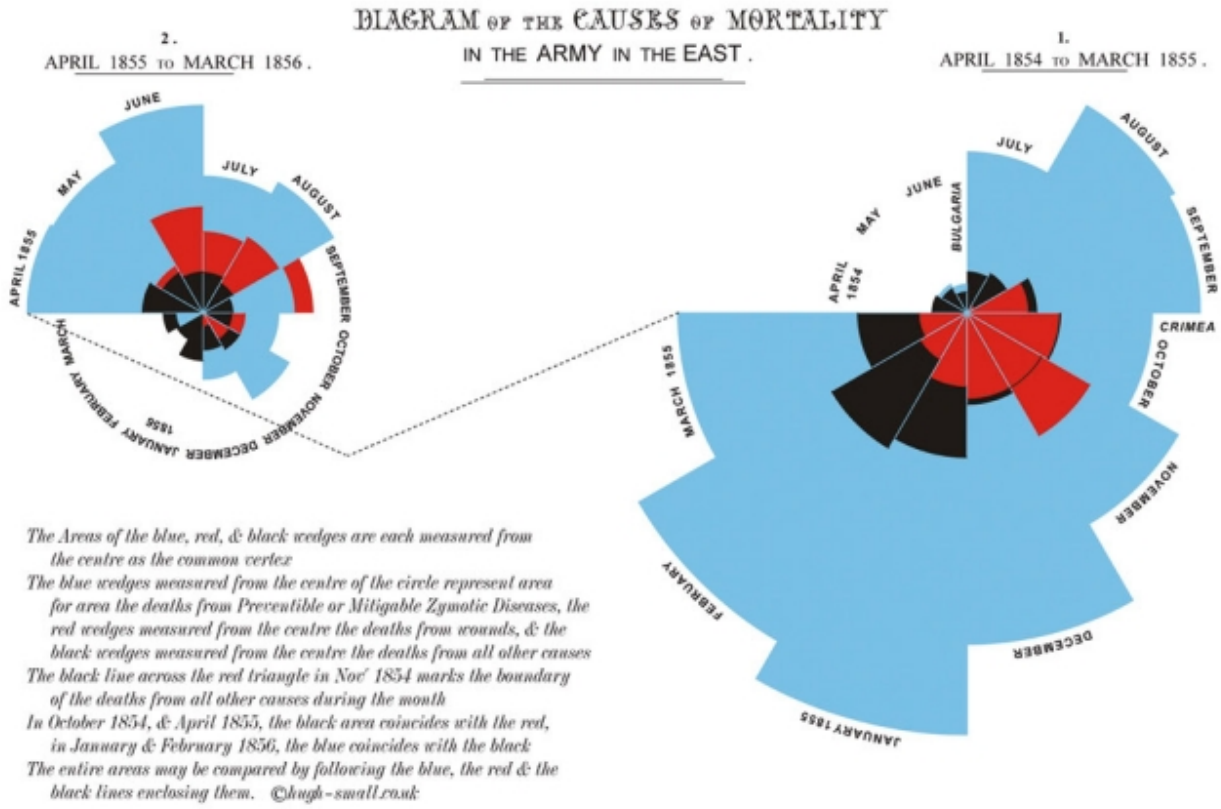
5 Data visualization using Python

5.1 Introduction

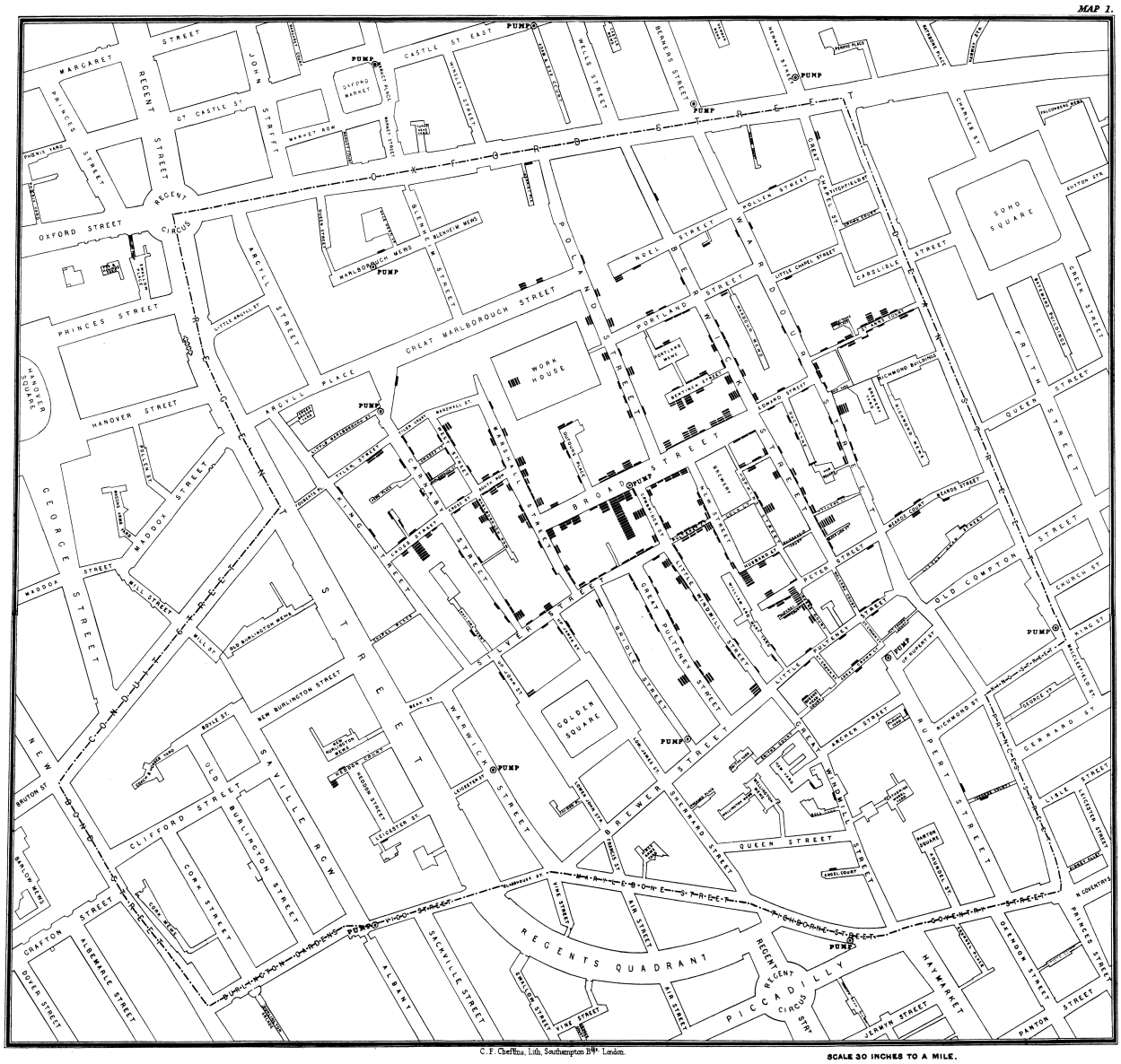
Data visualization is a basic task in data exploration and understanding. Humans are mainly visual creatures, and data visualization provides an opportunity to enhance communication of the story within the data. Often we find that data and the data-generating process is complex, and a visual representation of the data and our innate ability at pattern recognition can help reveal the complexities in a cognitively accessible way.

5.1.1 An example gallery

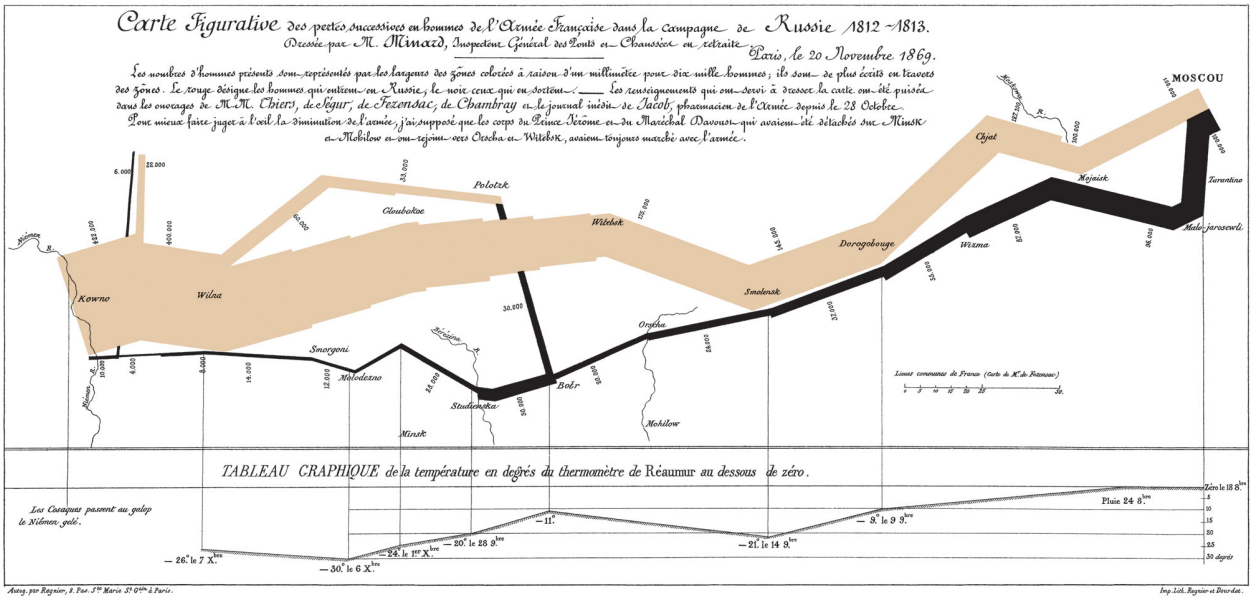
Data visualization has a long and storied history, from Florence Nightangle onwards. Dr. Nightangle was a pioneer in data visualization and developed the *rose plot* to represent causes of death in hospitals during the Crimean War.



John Snow, in 1854, famously visualized the cholera outbreak in London, which showed the geographic proximity of cholera prevalence with particular water wells.



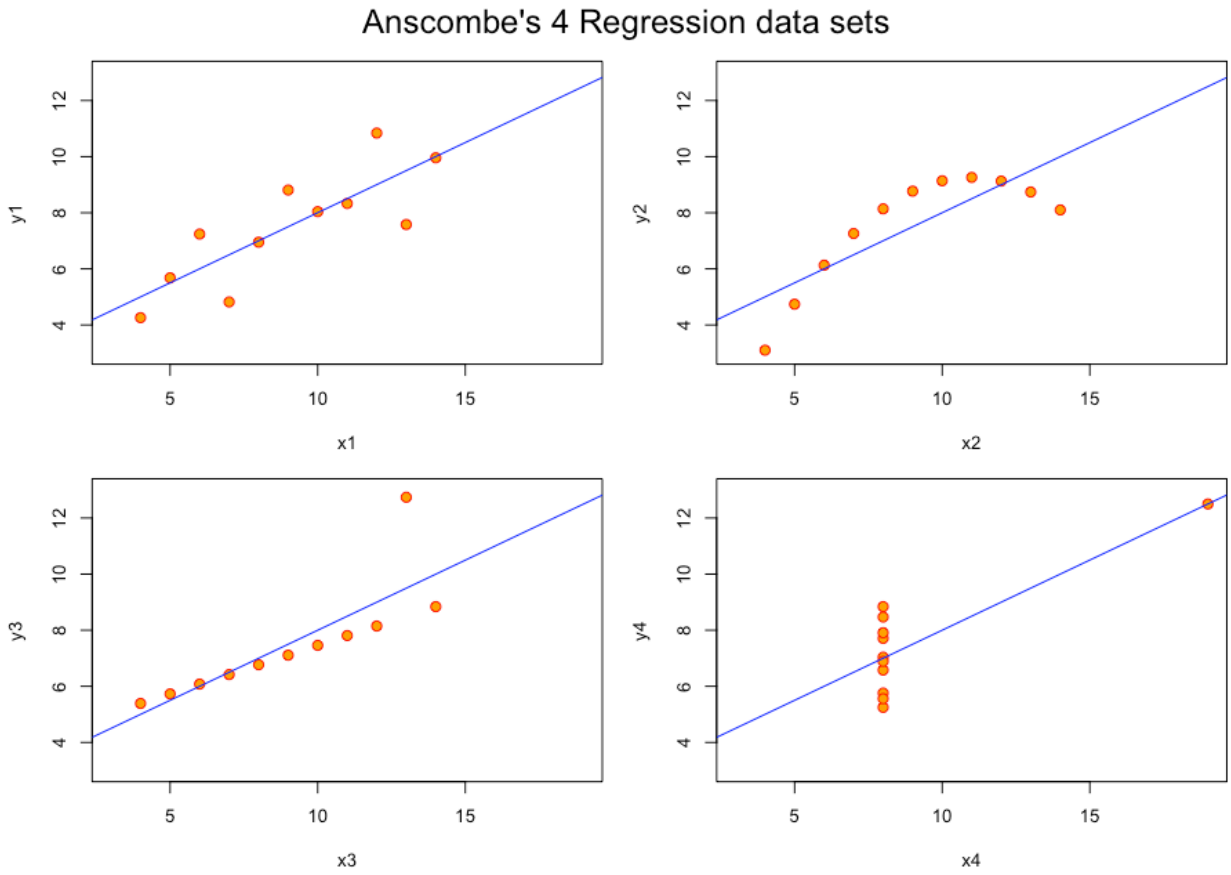
In one of the more famous visualizations, considered by many to be an optimal use of display ink and space, Minard visualized Napoleon's disastrous campaign to Russia



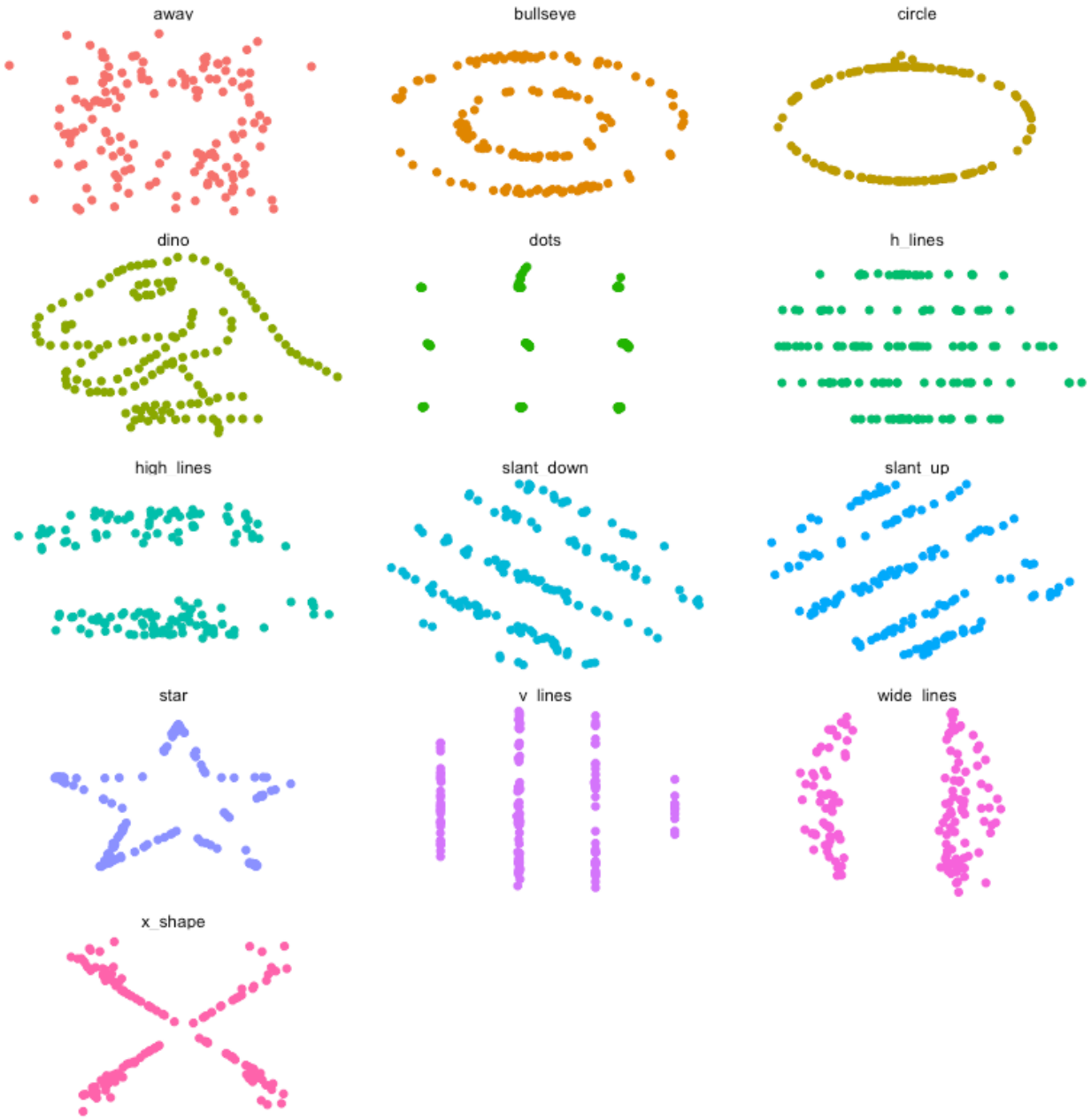
In more recent times, an employee at Facebook visualized all connections between users across the world, which clearly showed geographical associations with particular countries and regions.

5.1.2 Why visualize data?

We often rely on numerical summaries to help understand and distinguish datasets. In 1973, Anscombe published an influential set of 4 datasets, each with two variables and with the means, variances and correlations being identical. When you graphed these data, the differences in the datasets were clearly visible. This set is popularly known as Anscombe's quartet.



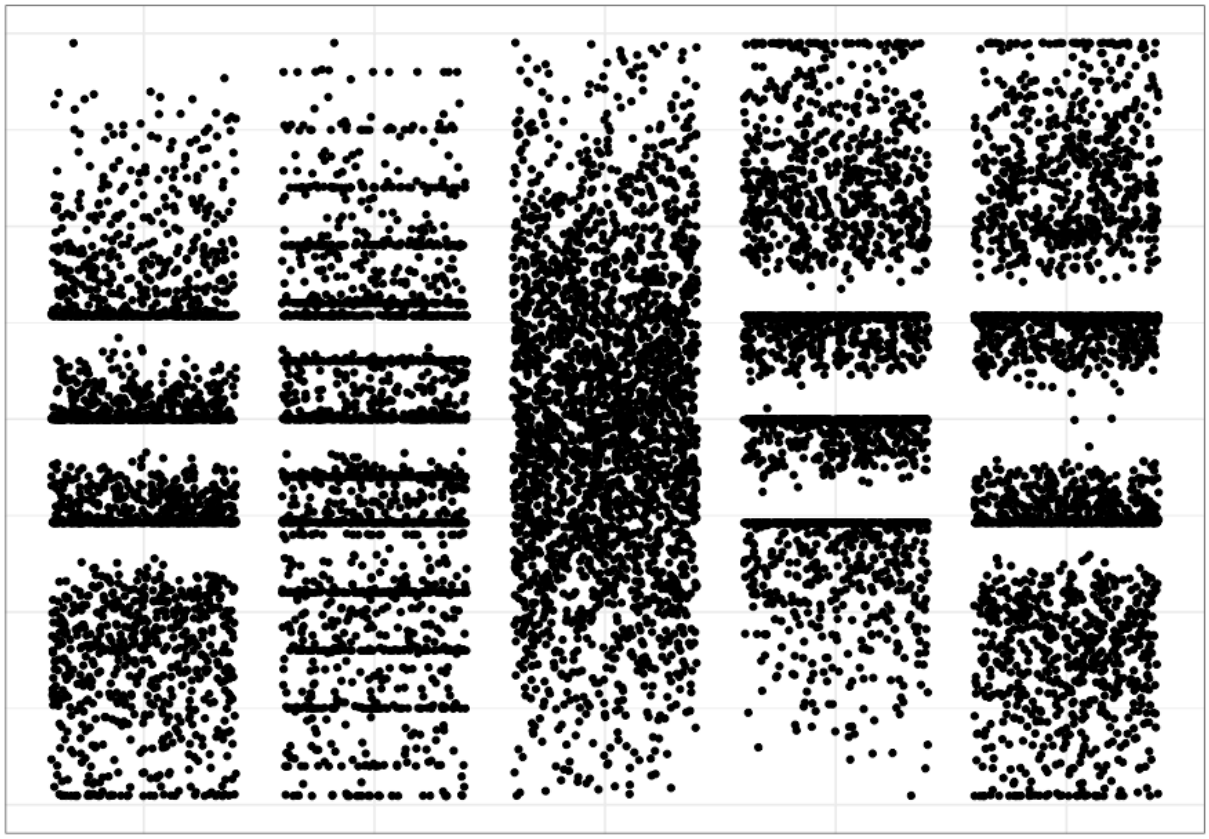
A more recent experiment in data construction by Matejka and Fitzmaurice (2017) started with a representation of a dinosaur and created 10 more bivariate datasets which all shared the same univariate means and variances and the same pairwise correlations.



These examples clarify the need for visualization to better understand relationships between variables.

Even when using statistical visualization techniques, one has to be careful. Not all visualizations can discriminate between statistical characteristics. This was also explored by Matejka and Fitzmaurice.

Strip plot



5.1.3 Conceptual ideas

5.1.3.1 Begin with the consumer in mind

- You have a deep understanding of the data you're presenting
- The person seeing the visualization **doesn't**
- Develop simpler visualizations first that are easier to explain

5.1.3.2 Tell a story

- Make sure the graphic is clear
- Make sure the main point you want to make "pops"

5.1.3.3 A matter of perception

- Color (including awareness of color deficiencies)
- Shape
- Fonts

5.1.3.4 Some principles

1. Data-ink ratio
2. No mental gymnastics
    1. The graphic should be self-evident
    2. Context should be clear
3. Is a graph the wrong choice?
4. Focus on the consumer

    See my slides for some more opinionated ideas

5.2 Plotting in Python

Let's take a very quick tour before we get into the weeds. We'll use the mtcars dataset as an exemplar dataset that we can import using `pandas`

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
sns.set_context('paper')
sns.set_style('white', {'font.family':'Futura', 'text.color':'1'})
```

```
sns.set_context('paper')
sns.set_style('white', {'font.family':'Futura Medium'})
```

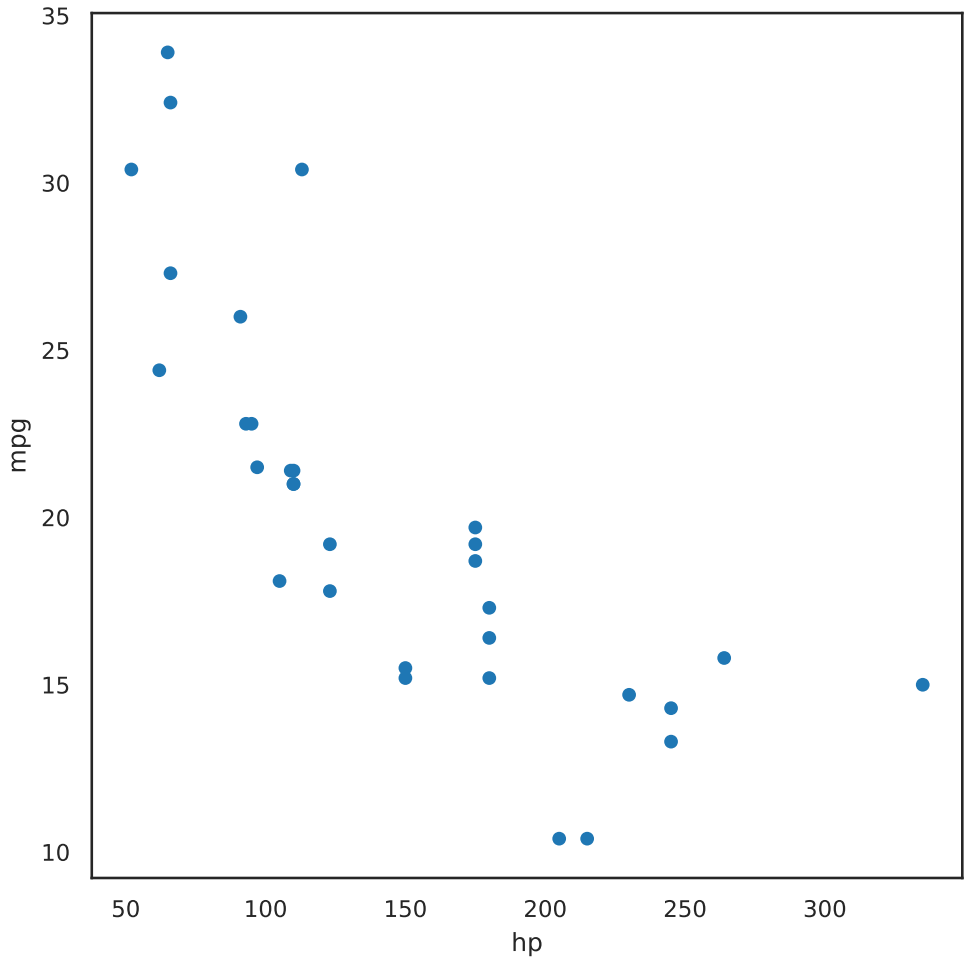
```
mtcars = pd.read_csv('data/mtcars.csv')
```

5.2.1 Static plots

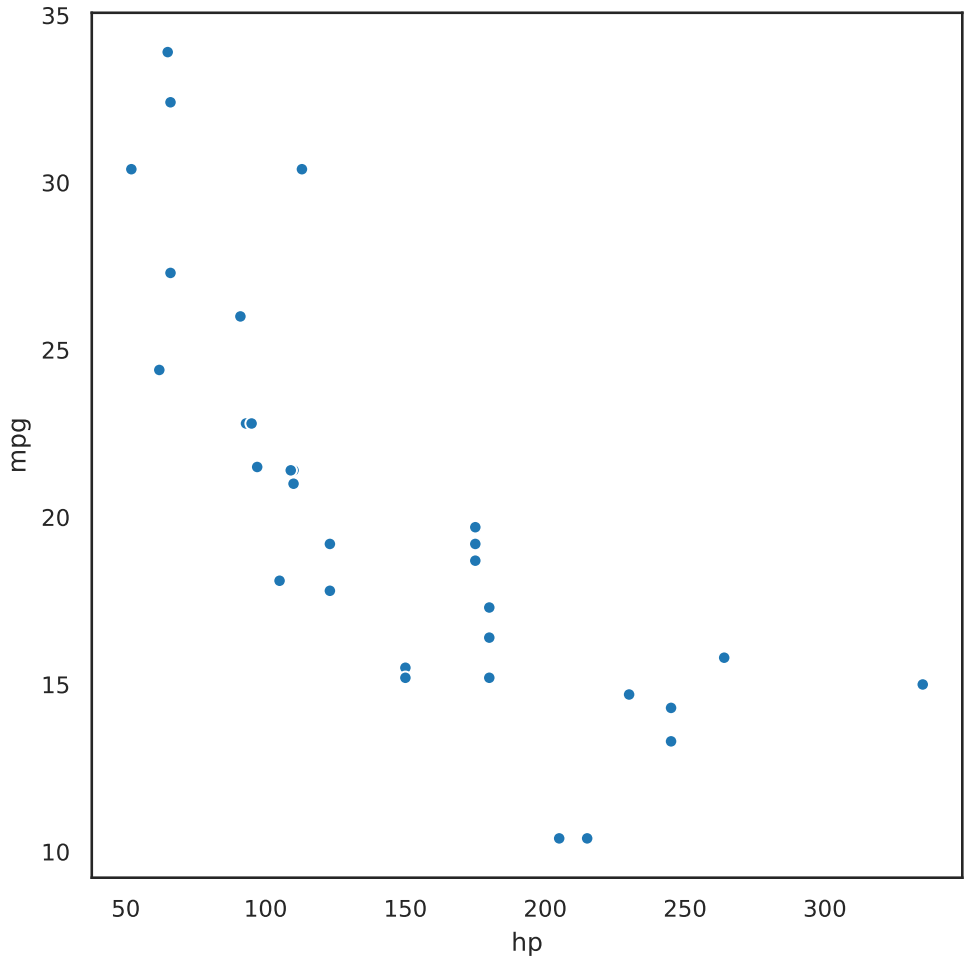
We will demonstrate plotting in what I'll call the `matplotlib` ecosystem. `matplotlib` is the venerable and powerful visualization package that was originally designed to emulate the Matlab plotting paradigm. It has since evolved and as become a bit more user-friendly. It is still quite granular and can facilitate a lot of custom plots once you become familiar with it. However, as a starting point, I think it's a bit much. We'll see a bit of what it can offer later.

We will consider two other options which are built on top of `matplotlib`, but are much more accessible. These are `pandas` and `seaborn`. The two packages have some different approaches, but both wrap `matplotlib` in higher-level code and decent choices so we don't need to get into the `matplotlib` trenches quite so much. We'll still call `matplotlib` in our code, since both these packages need it for some fine tuning. Both packages are also very much aligned to the `DataFrame` construct in `pandas`, so makes plotting a much more seamless experience.

```
mtcars.plot.scatter(x = 'hp', y = 'mpg');
plt.show()
# mtcars.plot(x = 'hp', y = 'mpg', kind = 'scatter');
```




```
sns.scatterplot(data = mtcars, x = 'hp', y = 'mpg');
plt.show()
```



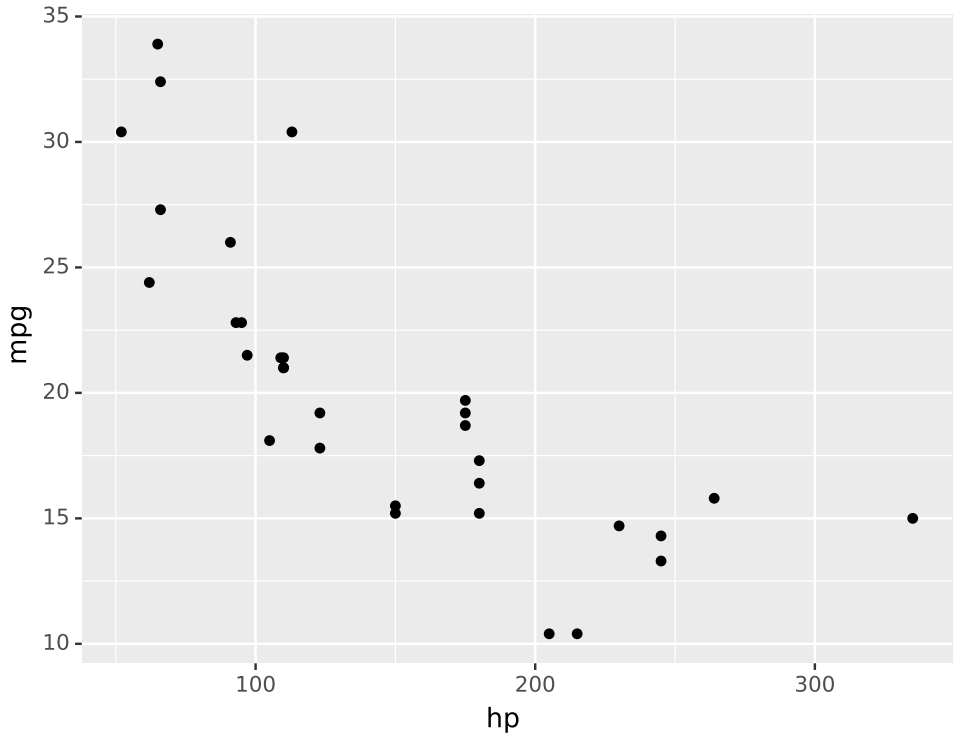
There are of course some other choices based on your background and preferences. For static plots, there are a couple of emulators of the popular R package `ggplot2`. These are `plotnine` and `ggplot`. `plotnine` seems a bit more developed and uses the `ggplot2` semantics of aesthetics and layers, with almost identical code syntax.

You can install `plotnine` using `conda`:

```
conda install -c conda-forge plotnine
```

```
from plotnine import *

(ggplot(mtcars) +
  aes(x = 'hp', y = 'mpg') +
  geom_point())
```

```
<ggplot: (329492989)>
```



5.2.2 Dynamic or interactive plots

There are several Python packages that wrap around Javascript plotting libraries that are so popular in web-based graphics like D3 and Vega. Three that deserve mention are `plotly`, `bokeh`, and `altair`.

> If you actually want to experience the interactivity of the plots, please use the "Live notebooks" link in Canvas to run these notebooks. Otherwise, you can download the notebooks from the GitHub site and run them on your own computer.

`plotly` is a Python package developed by the company Plot.ly to interface with their interactive Javascript library either locally or via their web service. Plot.ly also develops an R package to interface with their products as well. It provides an intuitive syntax and ease of use, and is probably the more popular package for interactive graphics from both R and Python.

```
import plotly.express as px

fig = px.scatter(mtcars, x = 'hp', y = 'mpg')
fig.show()
```

`bokeh` is an interactive visualization package developed by Anaconda. It is quite powerful, but its code can be rather verbose and granular

```
from bokeh.plotting import figure, output_file
from bokeh.io import output_notebook, show
output_notebook()
p = figure()
p.xaxis.axis_label = 'Horsepower'
p.yaxis.axis_label = 'Miles per gallon'

p.circle(mtcars['hp'], mtcars['mpg'], size=10);

show(p)
```

`altair` that leverages ideas from Javascript plotting libraries and a distinctive code syntax that may appeal to some

```
import altair as alt

alt.Chart(mtcars).mark_point().encode(
    x='hp',
    y='mpg'
).interactive()
```

We won't focus on these dynamic packages in this workshop in the interests of time, but you can avail of several online resources for these.

Package	Resources
plotly	Fundamentals
bokeh	Tutorial
altair	Overview

5.3 Univariate plots

We will be introducing plotting and code from 3 modules: `matplotlib`, `seaborn` and `pandas`. As we go forth, you may ask the question, which one should I learn? Chris Moffitt has the following advice.

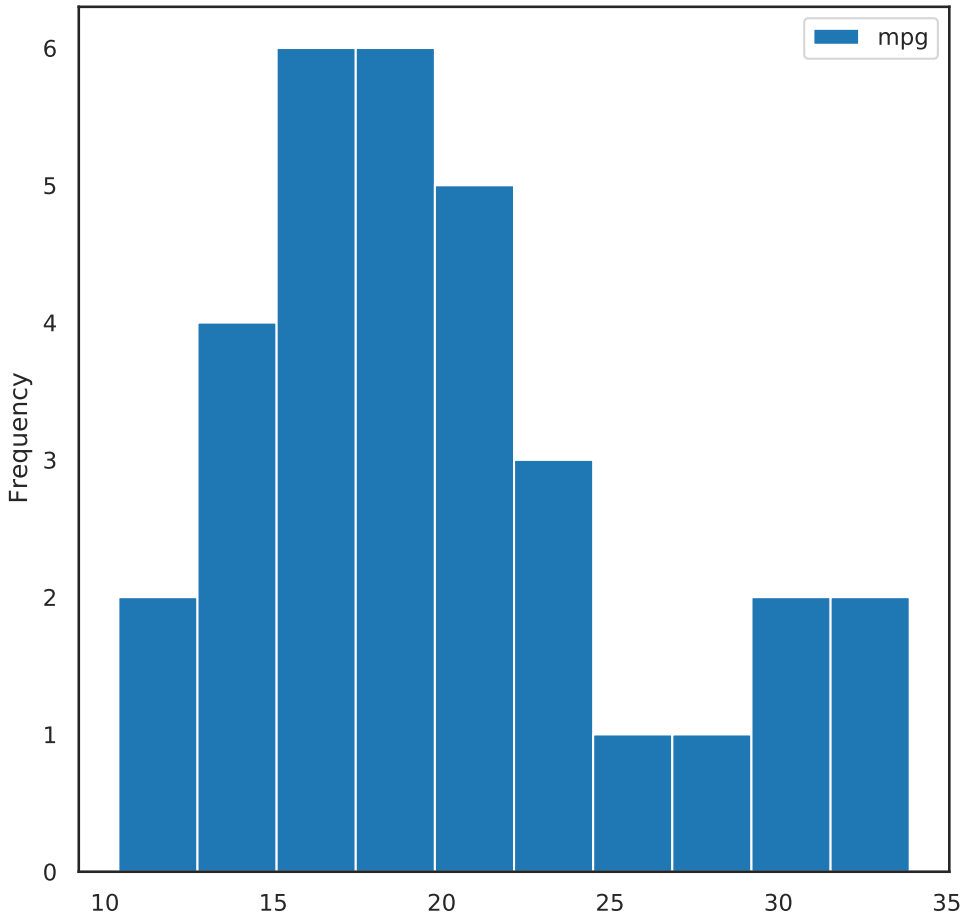
A pathway to learning (Chris Moffit)

1. Learn the basic matplotlib terminology, specifically what is a `Figure` and an `Axes` .
2. Always use the object-oriented interface. Get in the habit of using it from the start of your analysis. (*not really getting into this, but basically don't use the Matlab form I'll show at the end, if you don't have to*)
3. Start your visualizations with basic pandas plotting.
4. Use seaborn for the more complex statistical visualizations.
5. Use matplotlib to customize the pandas or seaborn visualization.

5.3.1 pandas

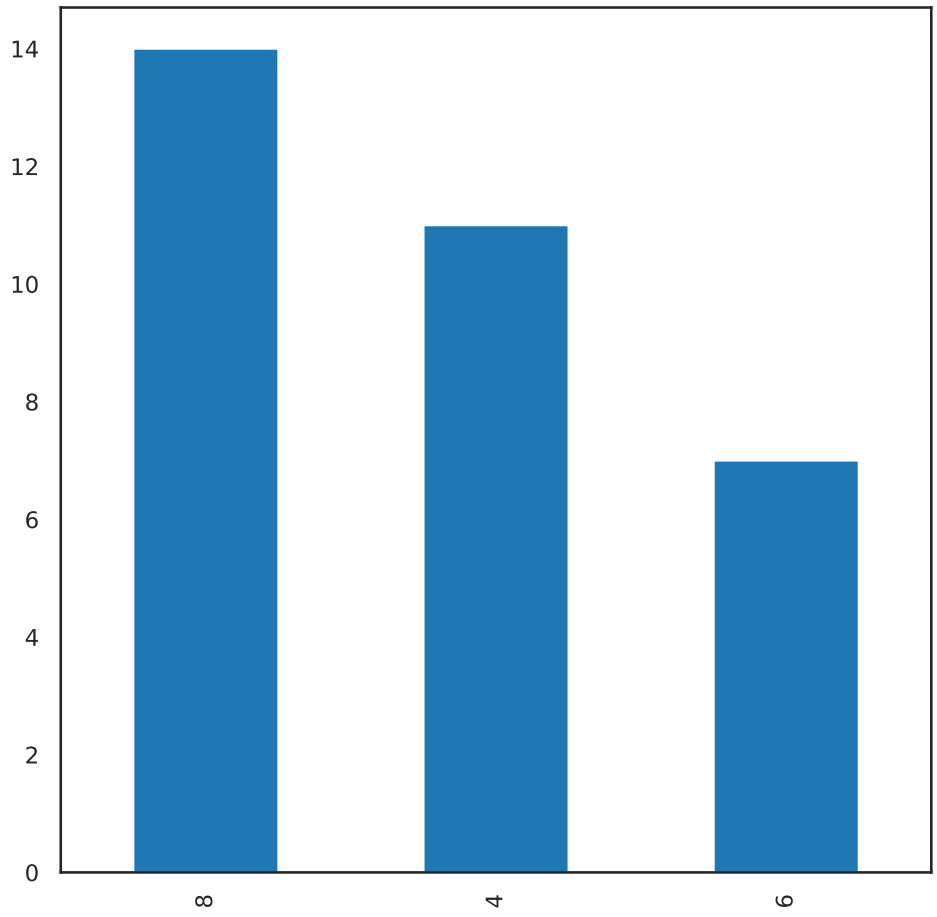
5.3.1.1 Histogram

```
mtcars.plot.hist(y = 'mpg');
plt.show()
# mtcars.plot(y = 'mpg', kind = 'hist')
#mtcars['mpg'].plot(kind = 'hist')
```



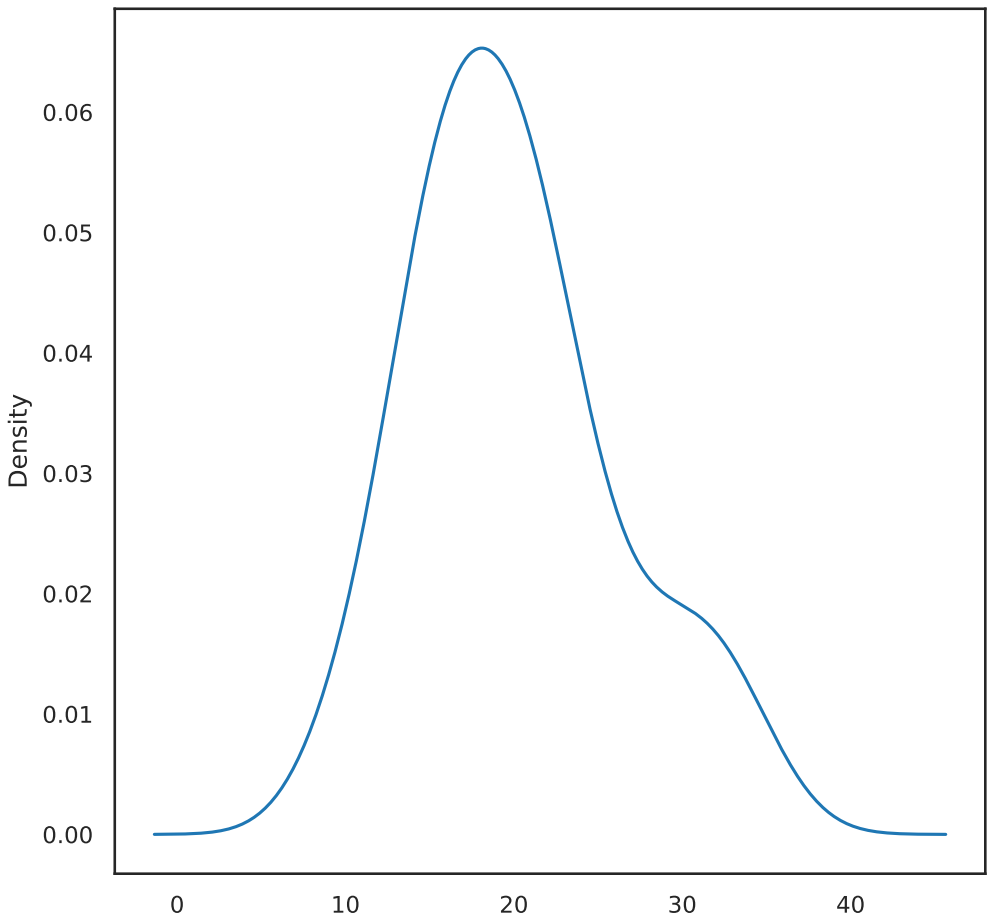
5.3.1.2 Bar plot

```
mtcars['cyl'].value_counts().plot.bar();
plt.show()
```



5.3.1.3 Density plot

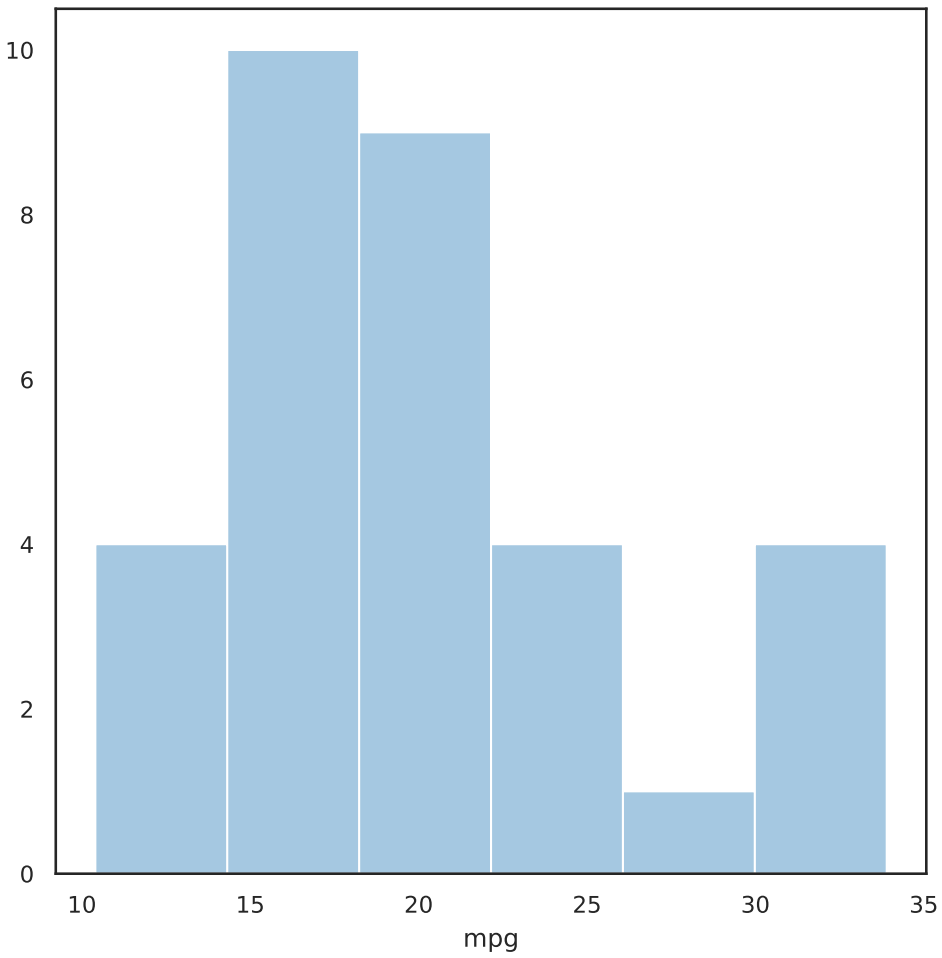
```
mtcars['mpg'].plot( kind = 'density');
plt.show()
```



5.3.2 seaborn

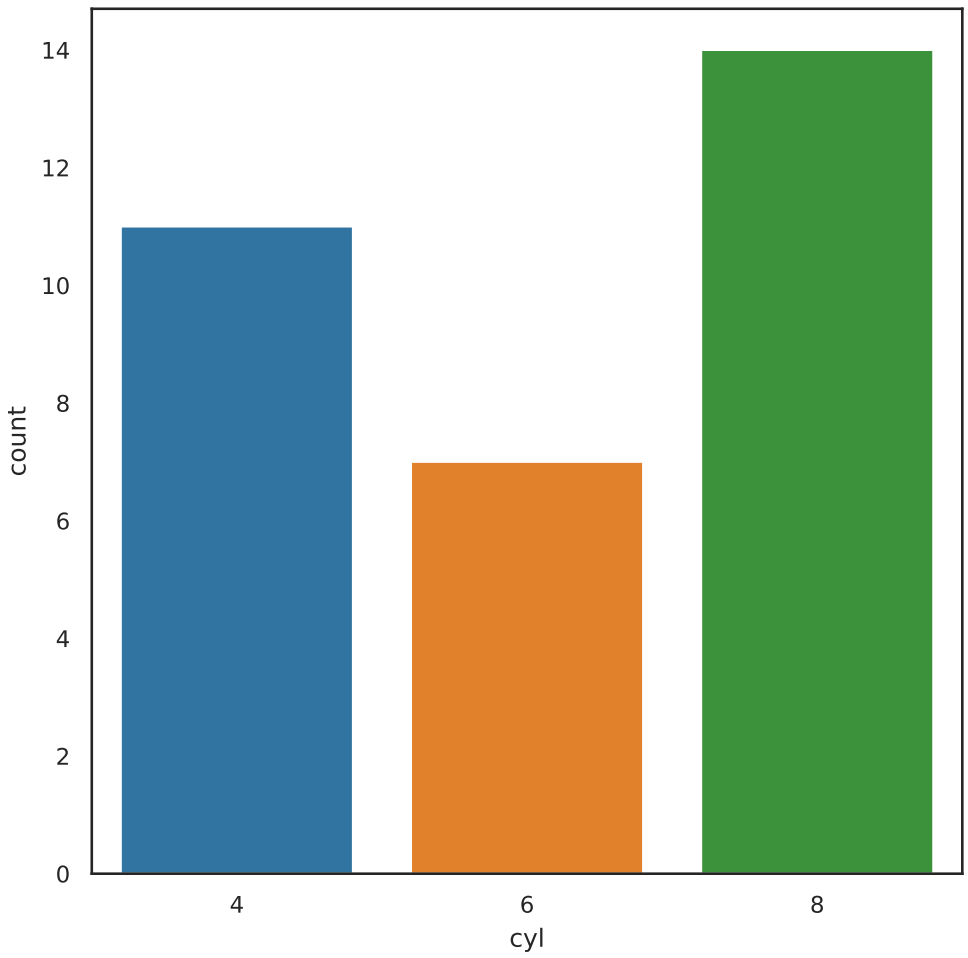
5.3.2.1 Histogram

```
ax = sns.distplot(mtcars['mpg'], kde=False);
plt.show()
```



5.3.2.2 Bar plot

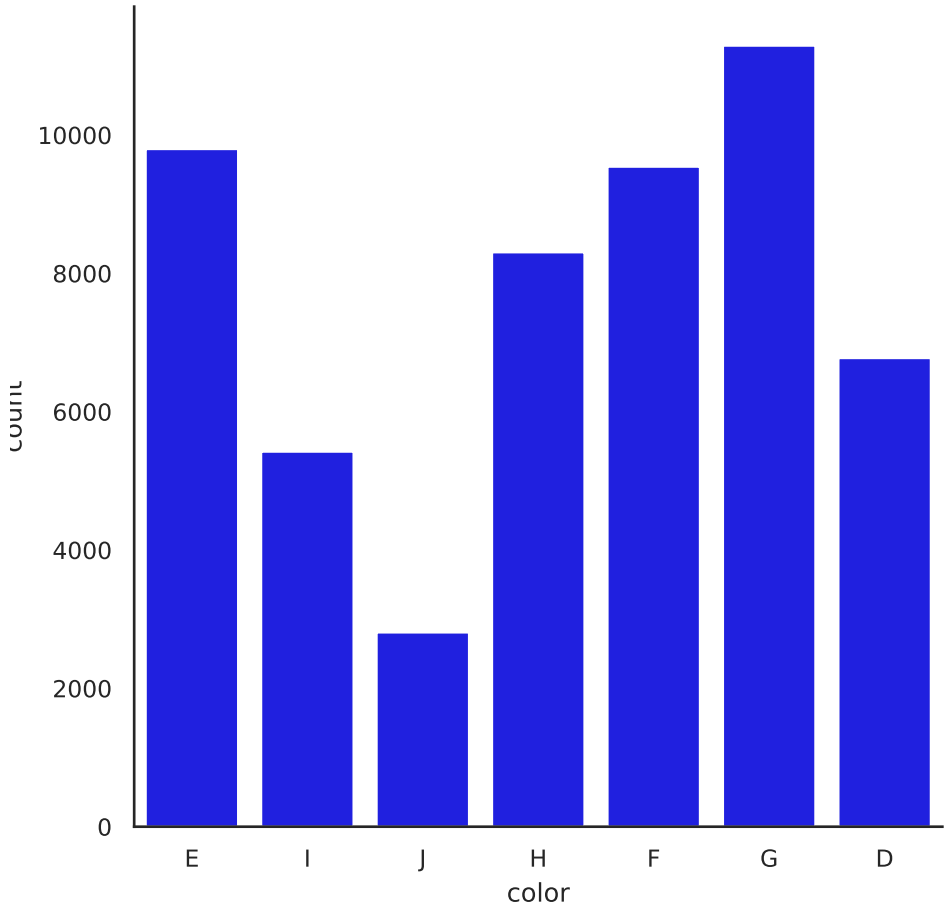
```
sns.countplot(data = mtcars, x = 'cyl');
plt.show()
```




```
diamonds = pd.read_csv('data/diamonds.csv.gz')
ordered_colors = ['E','F','G','H','I','J']
sns.catplot(data = diamonds, x = 'color', kind = 'count', color = 'blue');
```

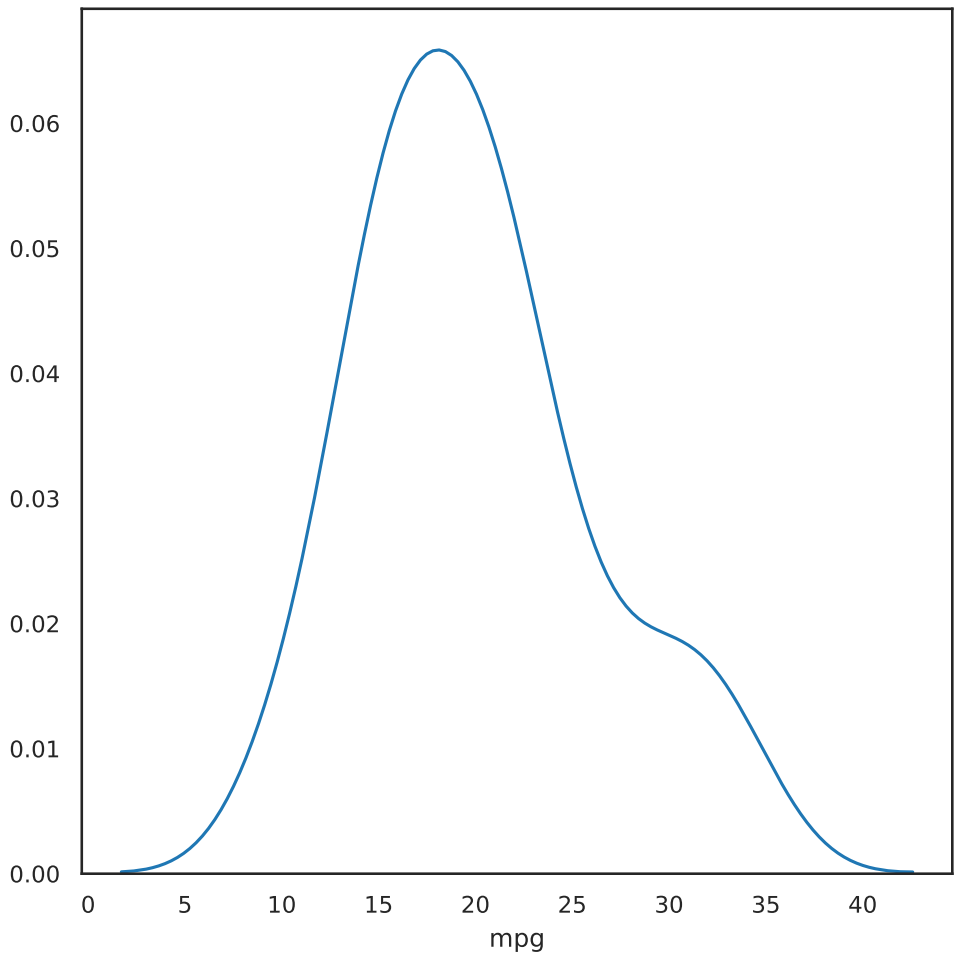
```
<seaborn.axisgrid.FacetGrid object at 0x139d3bd00>
```

```
plt.show()
```



5.3.2.3 Density plot

```
sns.distplot(mtcars['mpg'], hist=False);
plt.show()
```

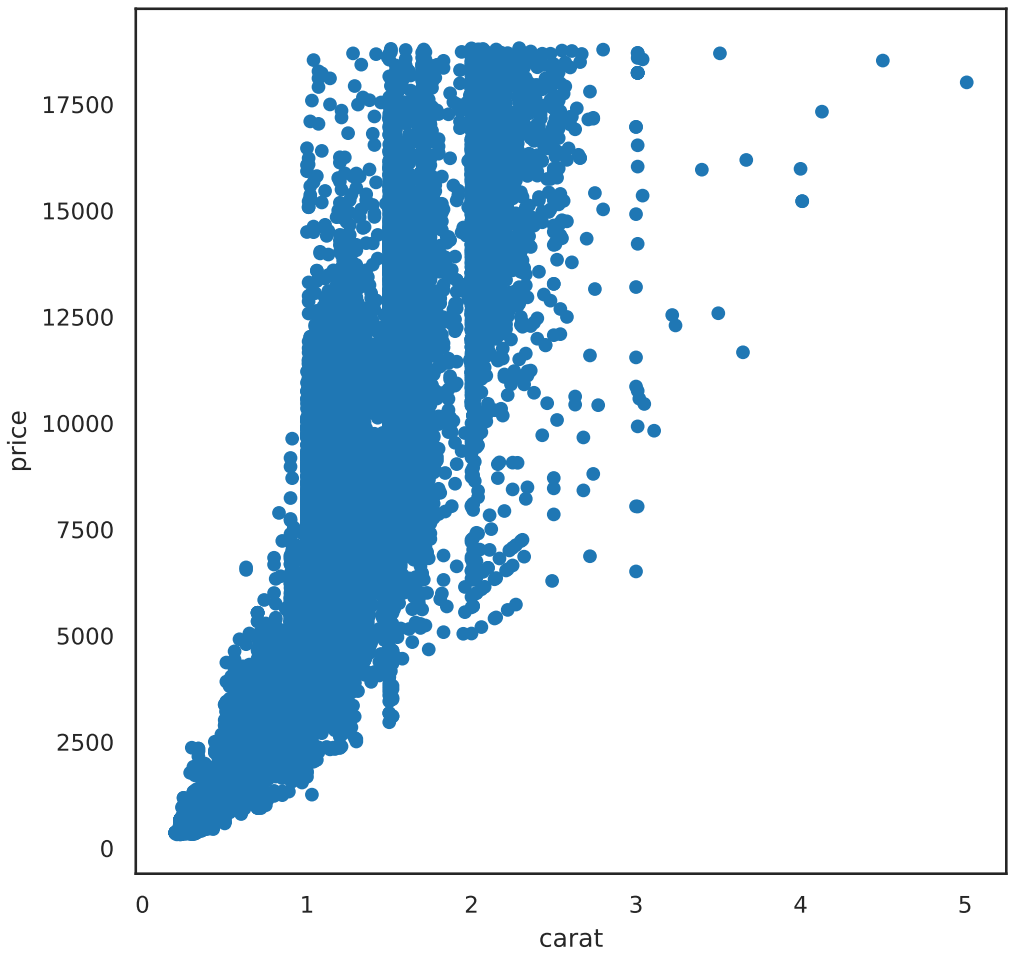


5.4 Bivariate plots

5.4.1 pandas

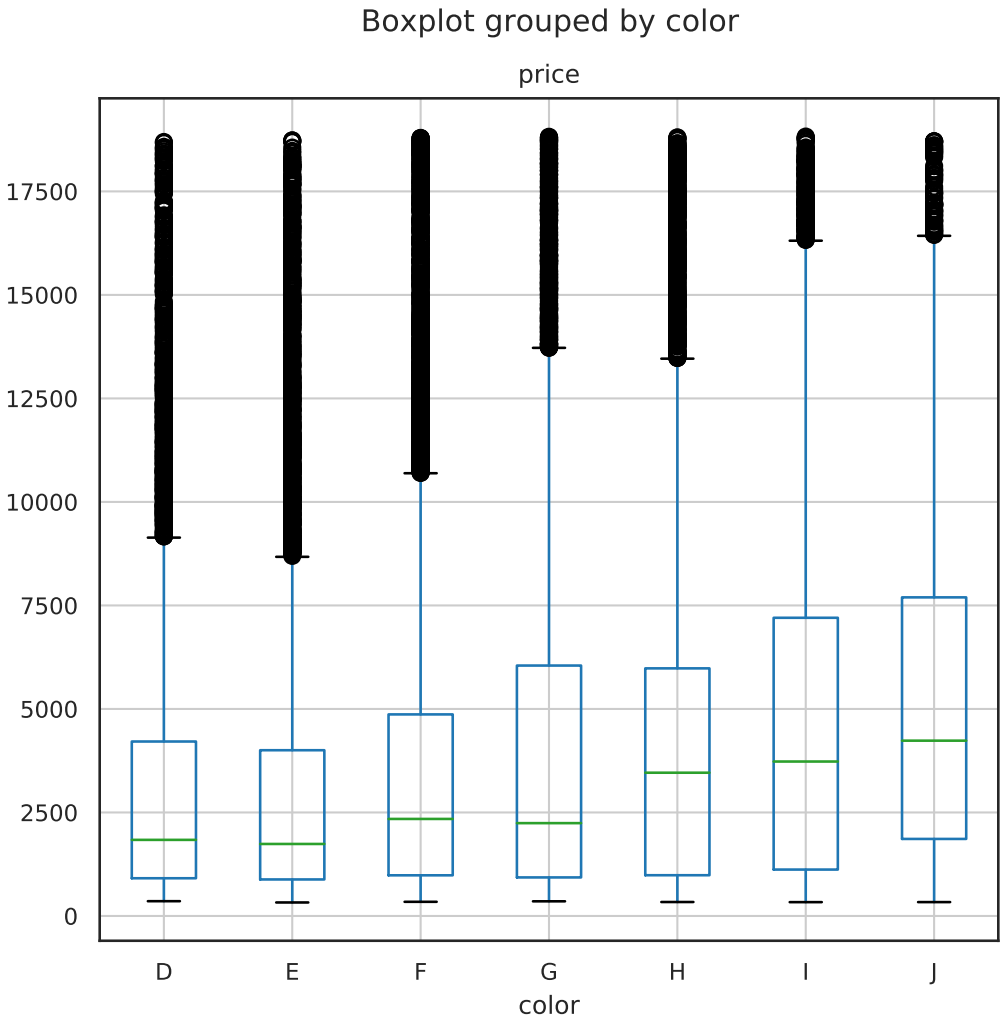
5.4.1.1 Scatter plot

```
diamonds = pd.read_csv('data/diamonds.csv.gz')
diamonds.plot(x = 'carat', y = 'price', kind = 'scatter');
plt.show()
```



5.4.1.2 Box plot

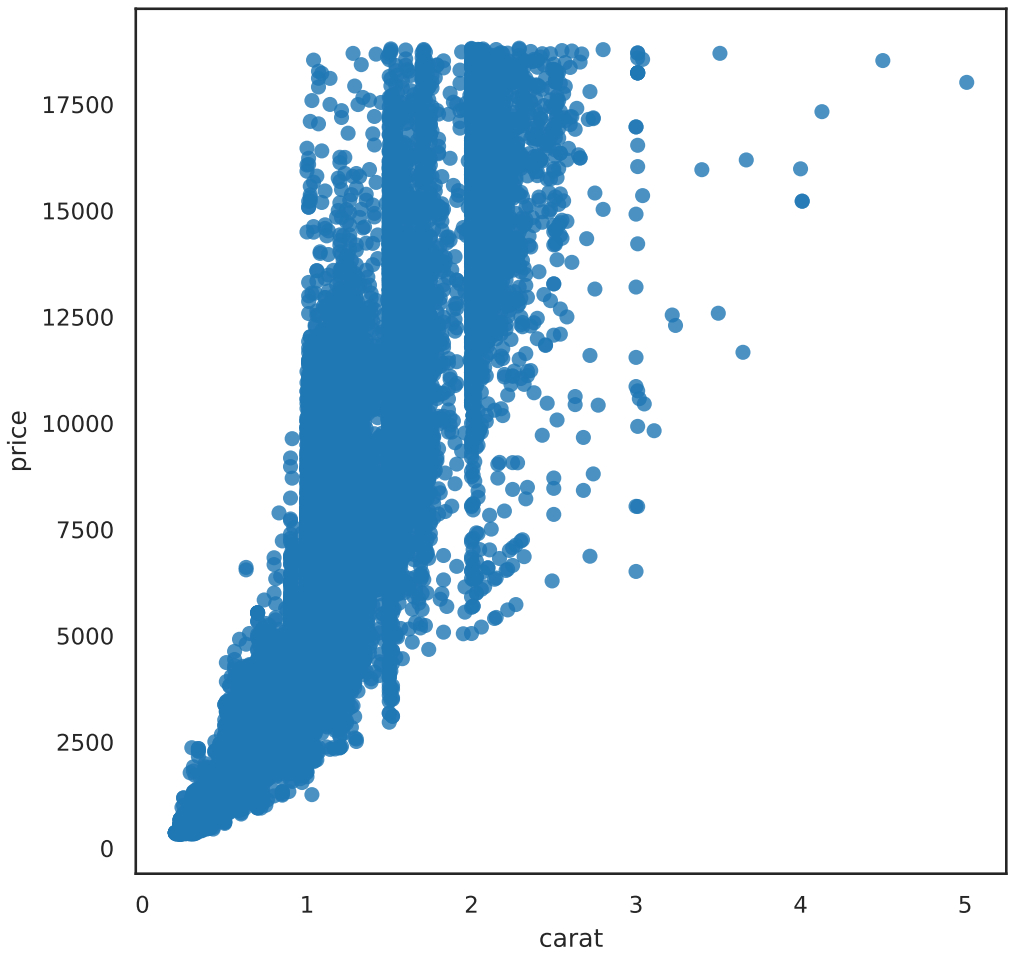
```
diamonds.boxplot(column = 'price', by = 'color');
plt.show()
```



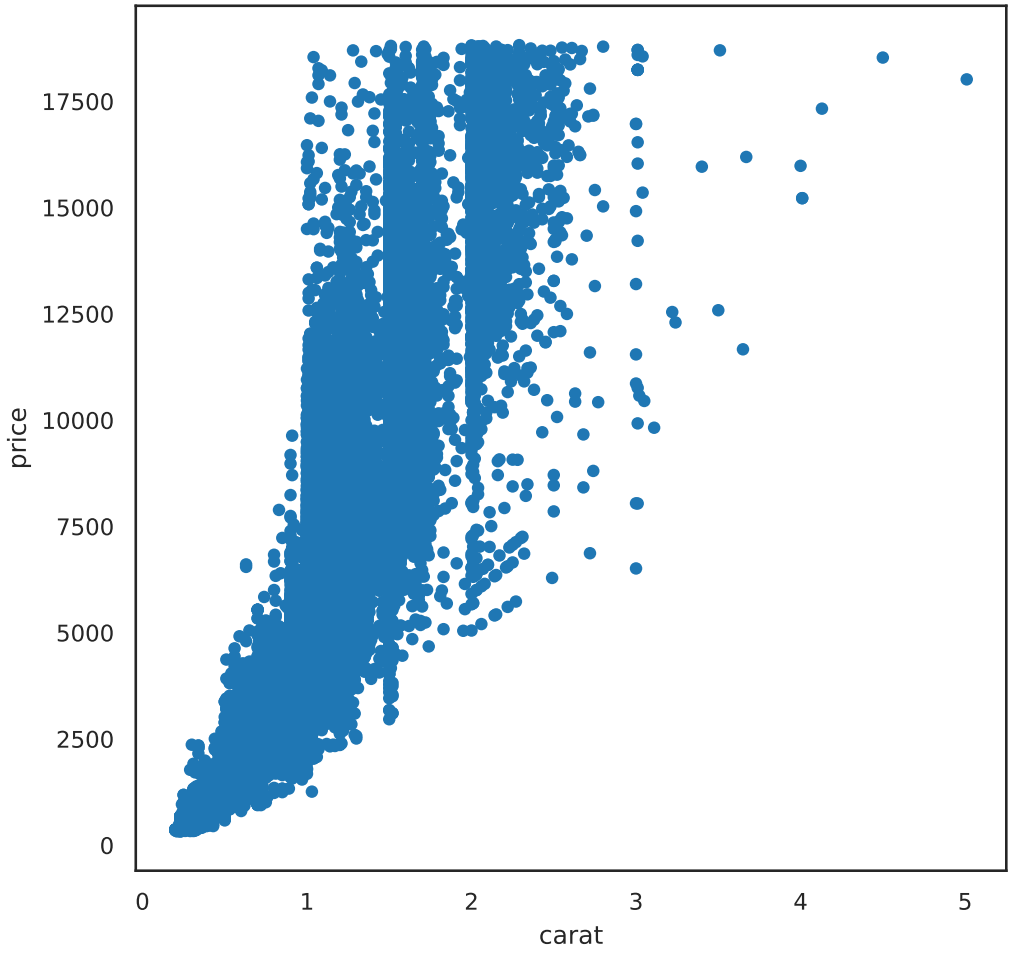
5.4.2 seaborn

5.4.2.1 Scatter plot

```
sns.regplot(data = diamonds, x = 'carat', y = 'price', fit_reg=False);
plt.show()
```




```
sns.scatterplot(data=diamonds, x = 'carat', y = 'price', linewidth=0);
# We set the linewidth to 0, otherwise the lines around the circles
# appear white and wash out the figure. Try with any positive
# value of linewidth
```

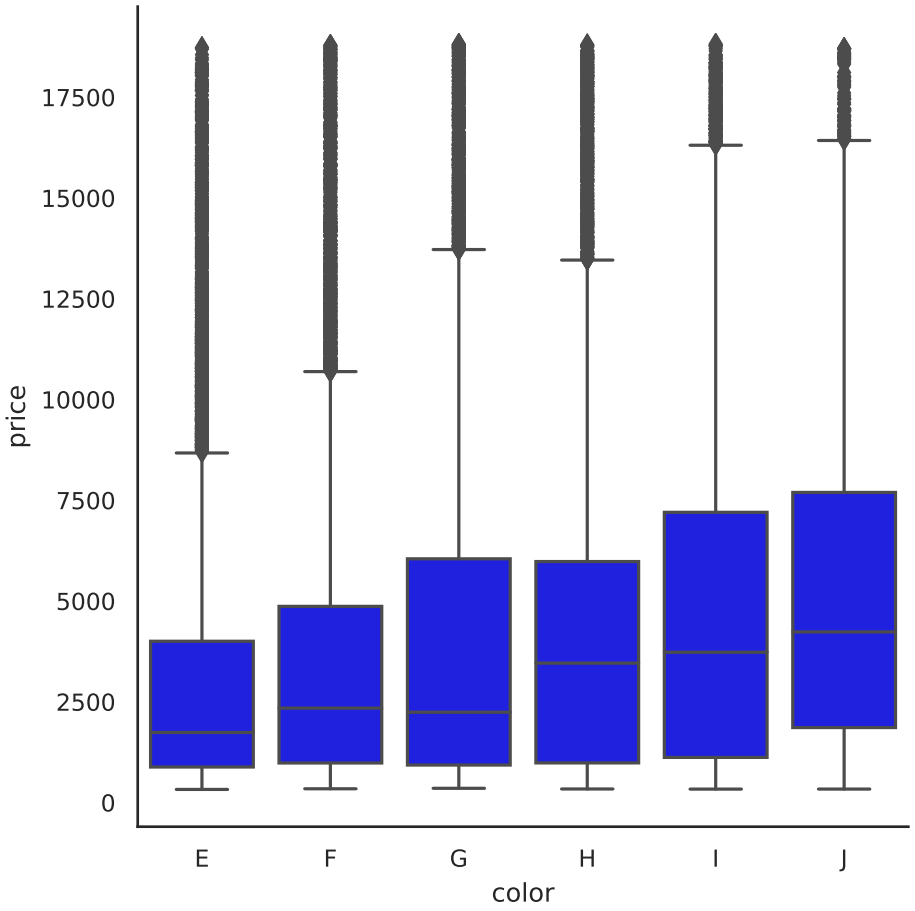


5.4.2.2 Box plot

```
ordered_color = ['E','F','G','H','I','J']
sns.catplot(data = diamonds, x = 'color', y = 'price',
            order = ordered_color, color = 'blue', kind = 'box');
```

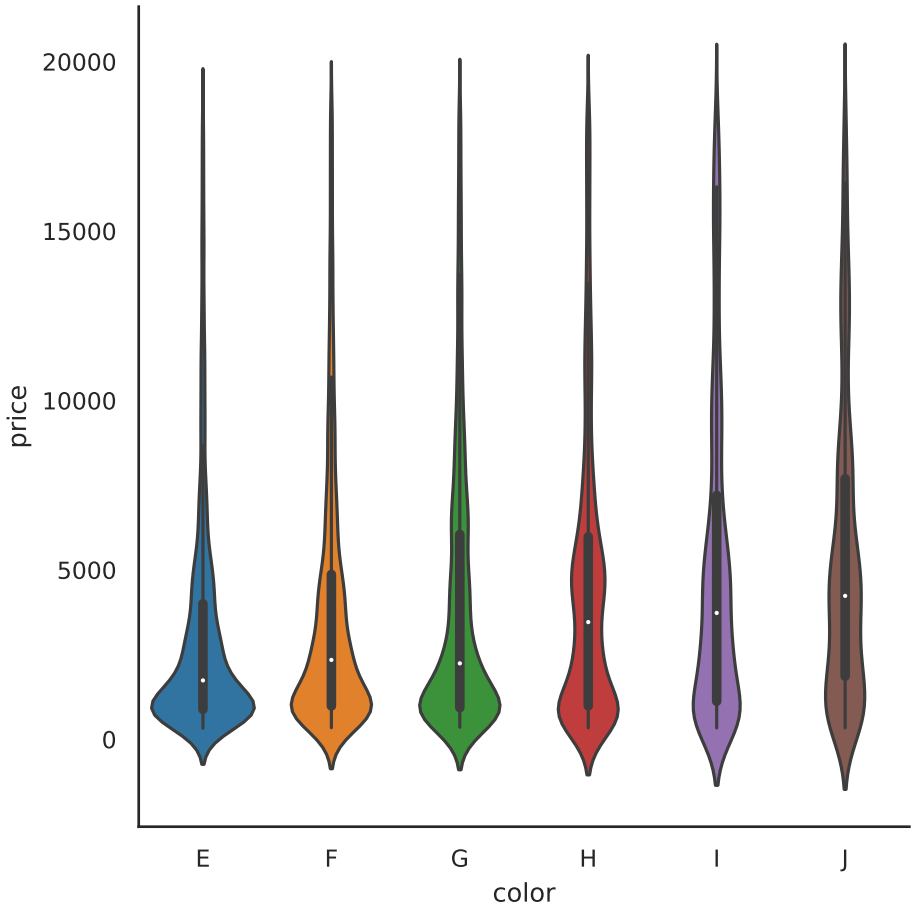
```
<seaborn.axisgrid.FacetGrid object at 0x139a11c10>
```

```
plt.show()
```



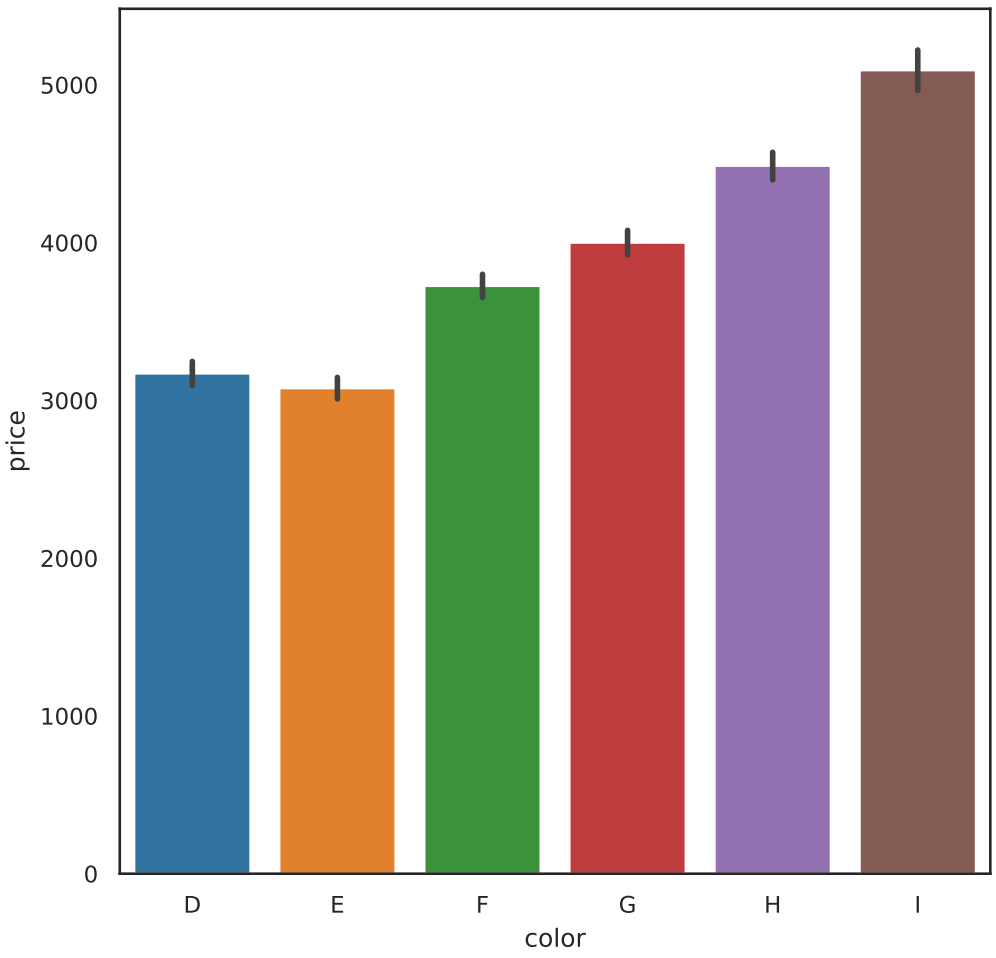
5.4.2.3 Violin plot

```
g = sns.catplot(data = diamonds, x = 'color', y = 'price',
                kind = 'violin', order = ordered_color);
plt.show()
```

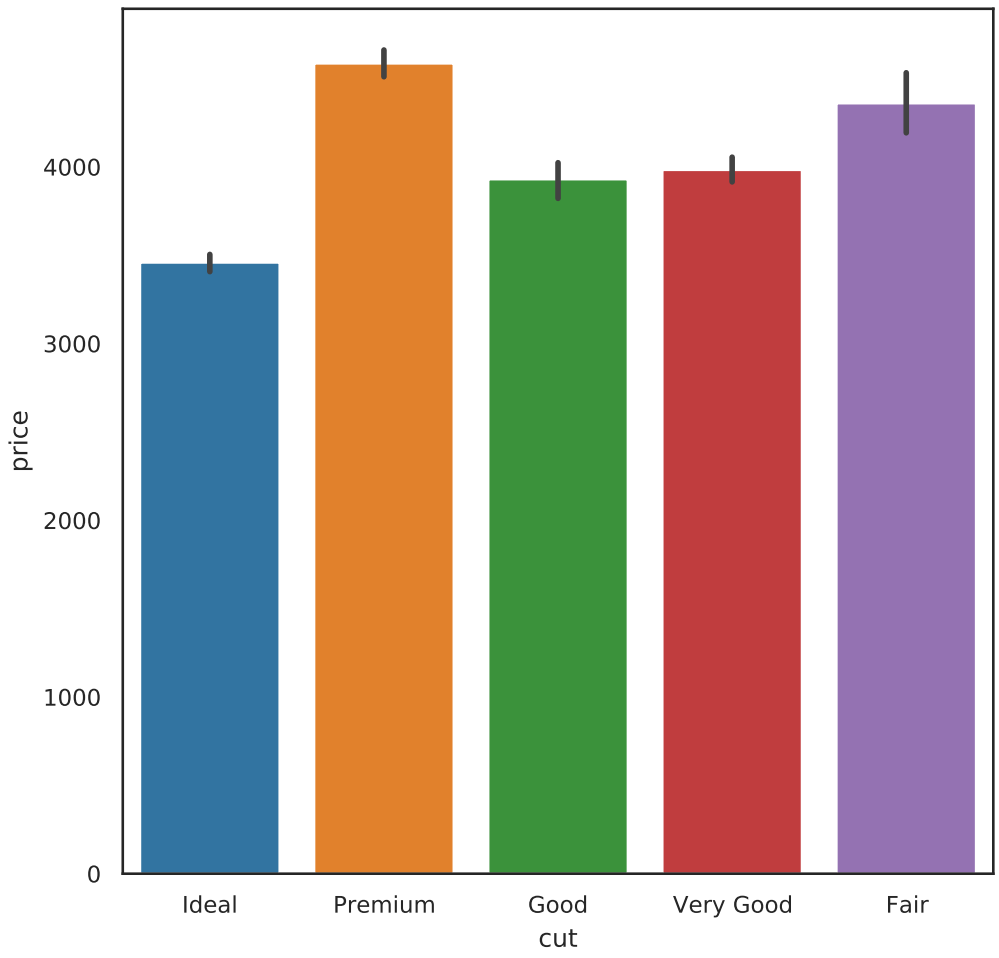


5.4.2.4 Barplot (categorical vs continuous)

```
ordered_colors = ['D','E','F','G','H','I']
sns.barplot(data = diamonds, x = 'color', y = 'price', order = ordered_colors);
plt.show()
```




```
sns.barplot(data = diamonds, x = 'cut', y = 'price');
plt.show()
```

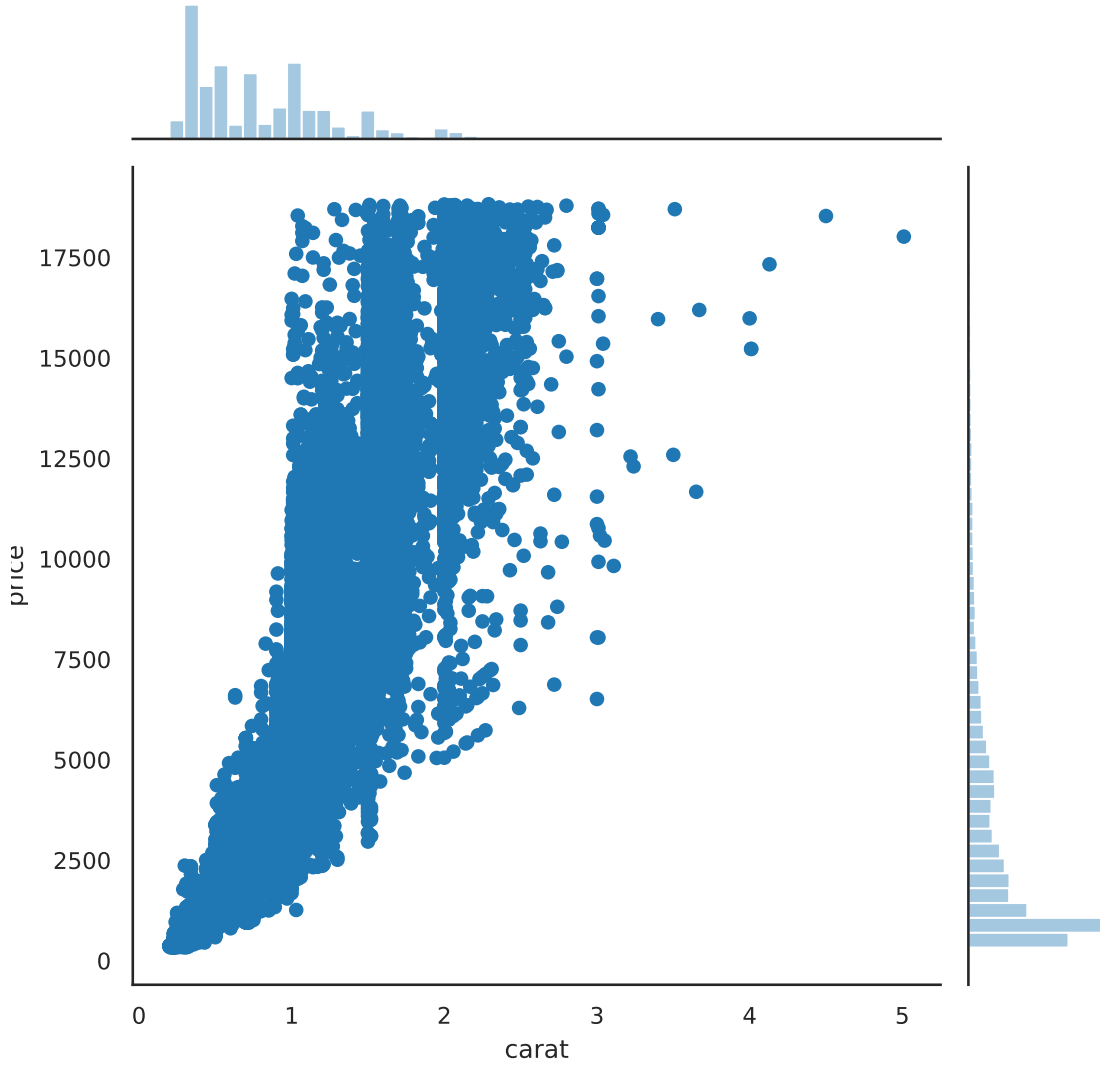


5.4.2.5 Joint plot

```
sns.jointplot(data = diamonds, x = 'carat', y = 'price');
```

```
<seaborn.axisgrid.JointGrid object at 0x13aa7a670>
```

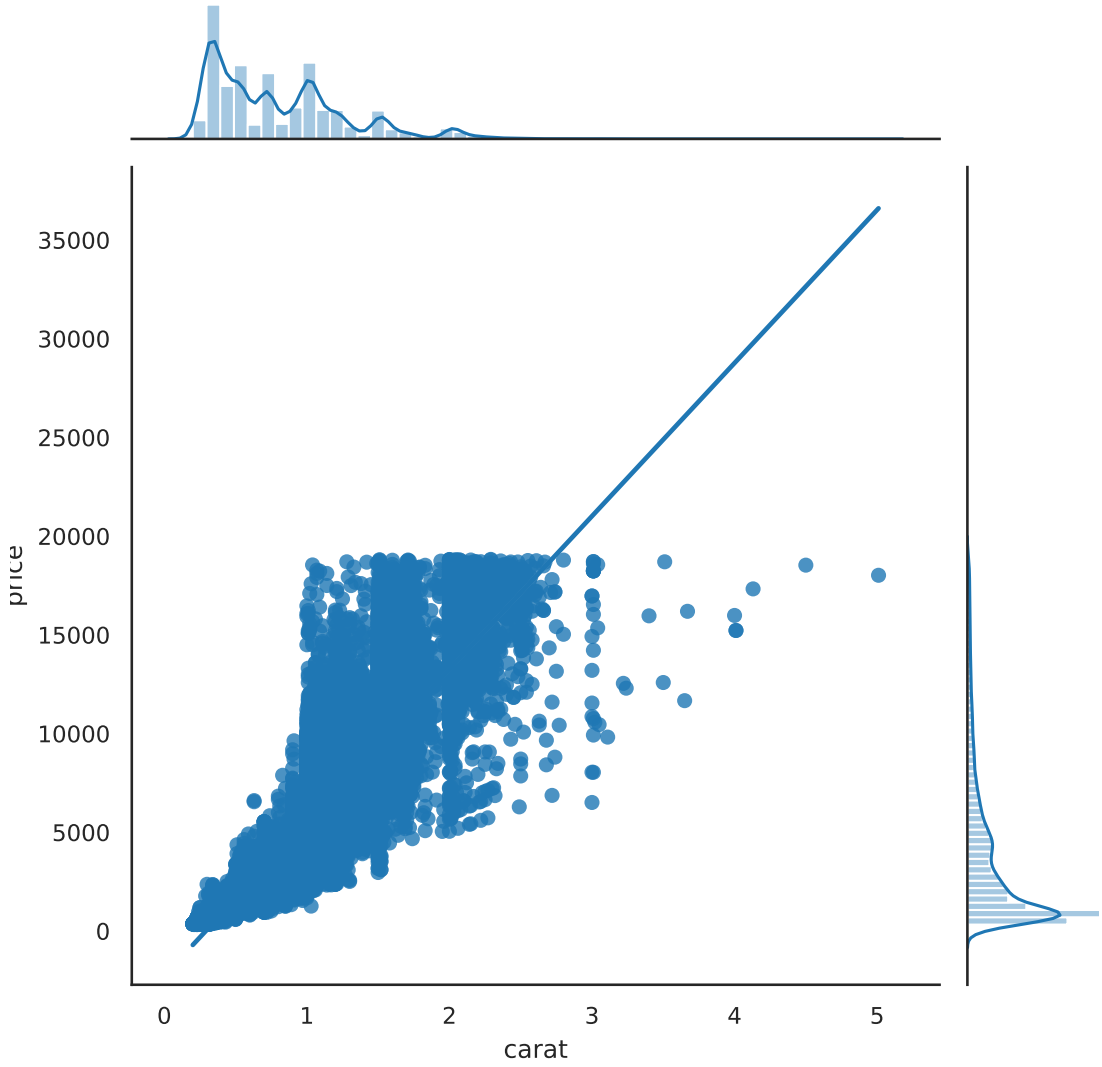
```
plt.show()
```




```
sns.jointplot(data = diamonds, x = 'carat', y = 'price', kind = 'reg');
```

```
<seaborn.axisgrid.JointGrid object at 0x1380d3c70>
```

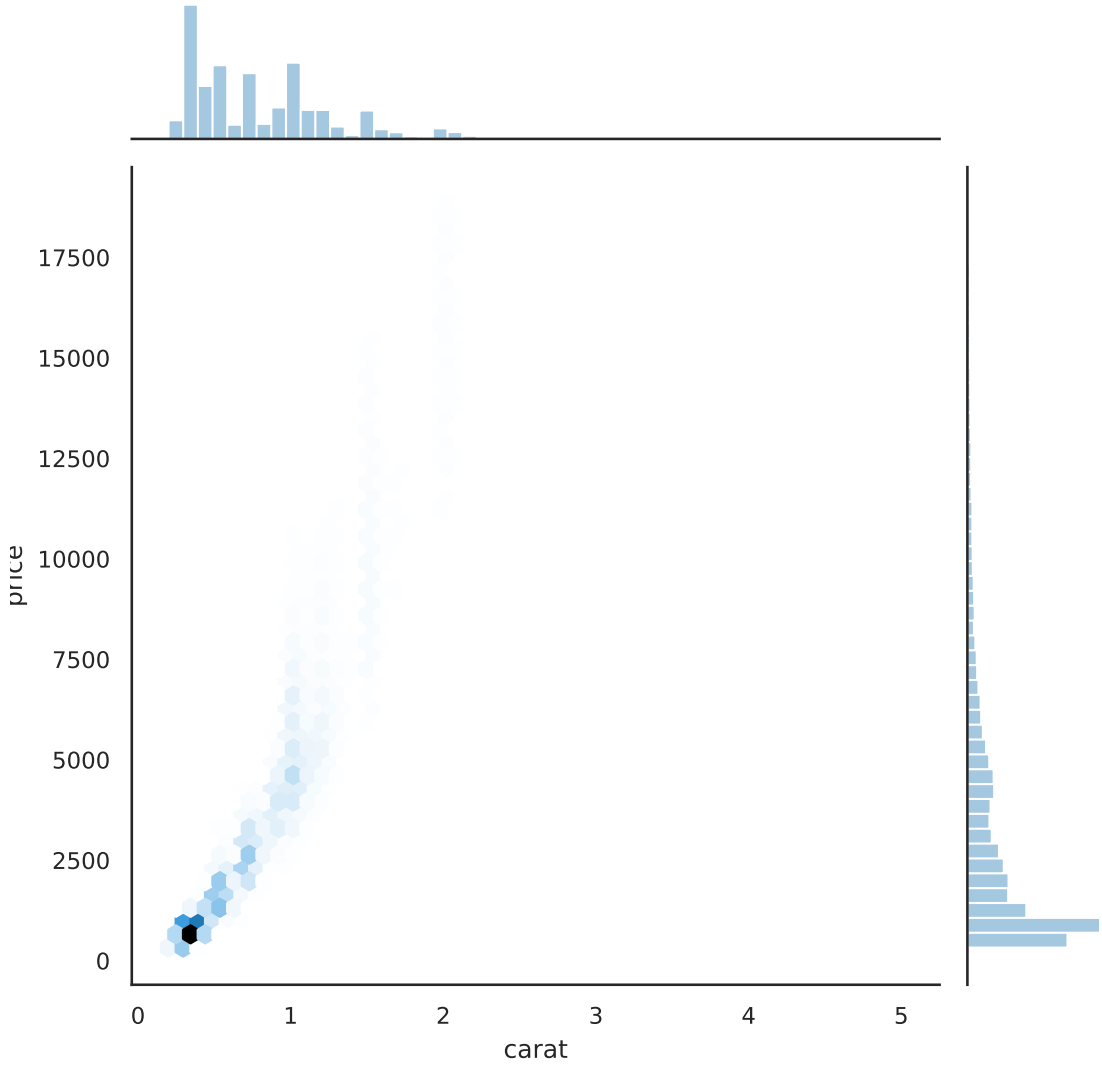
```
plt.show()
```




```
sns.jointplot(data = diamonds, x = 'carat', y = 'price', kind = 'hex');
```

```
<seaborn.axisgrid.JointGrid object at 0x138c92490>
```

```
plt.show()
```



5.5 Facets and multivariate data

The basic idea in this section is to see how we can visualize more than two variables at a time. We will see two strategies:

1. Put multiple graphs on the same frame, with each graph referring to a level of a 3rd variable
2. Create a grid of separate graphs, with each graph referring to a level of a 3rd variable

This strategy also can work any time we need to visualize the data corresponding to different levels of a variable, say by gender or race or country.

In this example we're going to start with 4 time series, labelled A, B, C, D.

```
ts = pd.read_csv('data/ts.csv')
ts.dt = pd.to_datetime(ts.dt) # convert this column to a datetime object
ts.head()
```

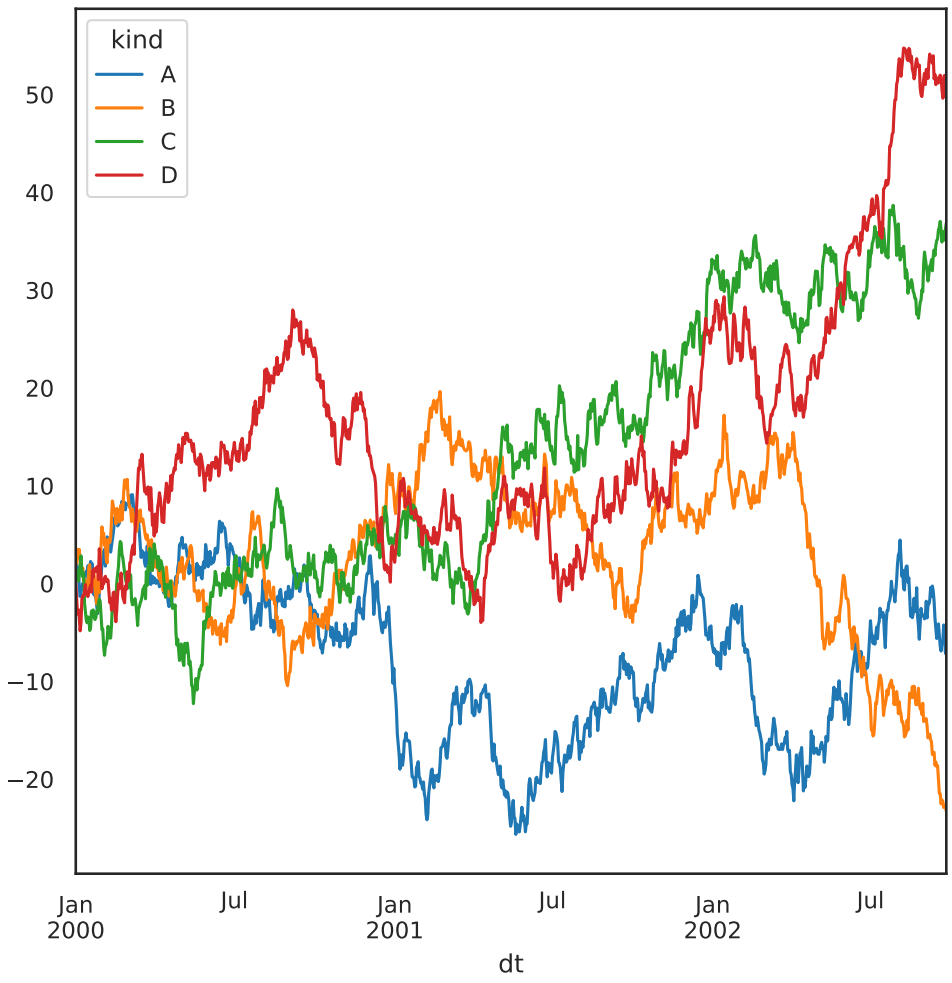
```
          dt kind     value
0 2000-01-01    A  1.442521
1 2000-01-02    A  1.981290
2 2000-01-03    A  1.586494
3 2000-01-04    A  1.378969
4 2000-01-05    A -0.277937
```

For one strategy we will employ, it is actually a bit easier to change this to a wide data form, using `pivot`.

```
dfp = ts.pivot(index = 'dt', columns = 'kind', values = 'value')
dfp.head()
```

```
kind                A         B         C         D
dt
2000-01-01   1.442521  1.808741  0.437415  0.096980
2000-01-02   1.981290  2.277020  0.706127 -1.523108
2000-01-03   1.586494  3.474392  1.358063 -3.100735
2000-01-04   1.378969  2.906132  0.262223 -2.660599
2000-01-05  -0.277937  3.489553  0.796743 -3.417402
```

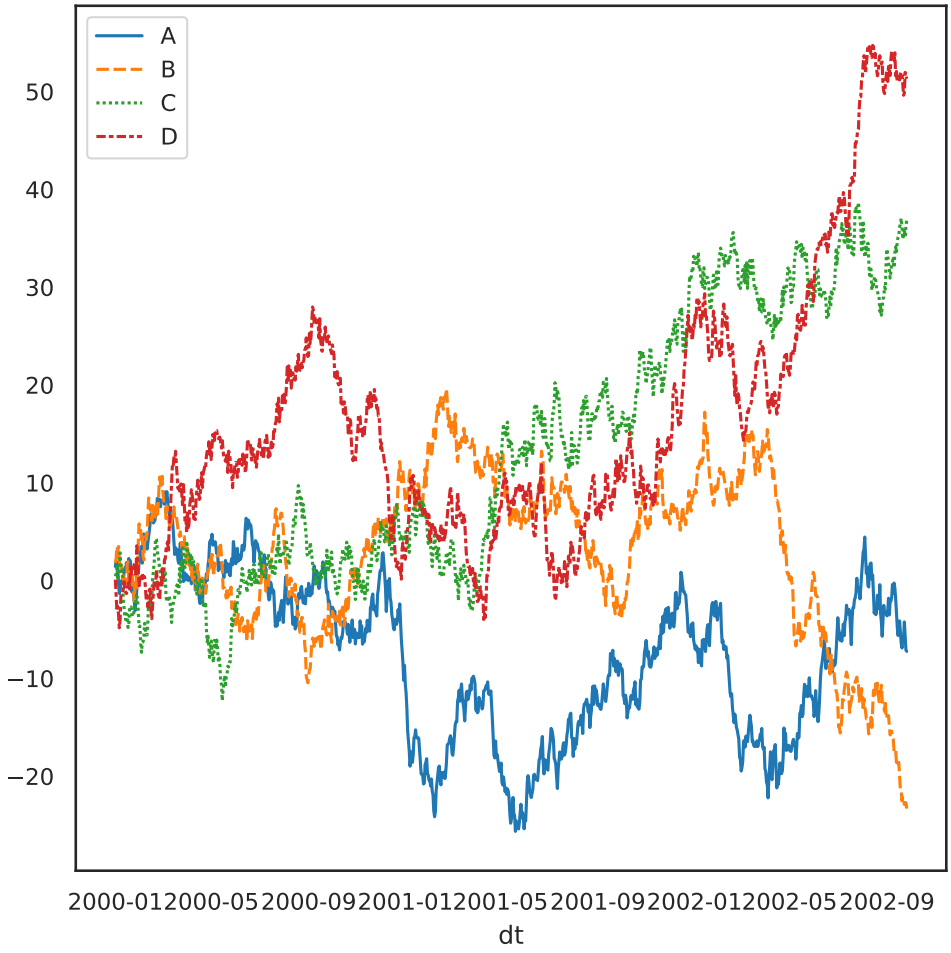
```
fig, ax = plt.subplots()
dfp.plot(ax=ax);
plt.show()
```



This creates 4 separate time series plots, one for each of the columns labeled A, B, C and D. The x-axis is determined by `dfp.index`, which during the pivoting operation, we deemed was the values of `dt` in the original data.

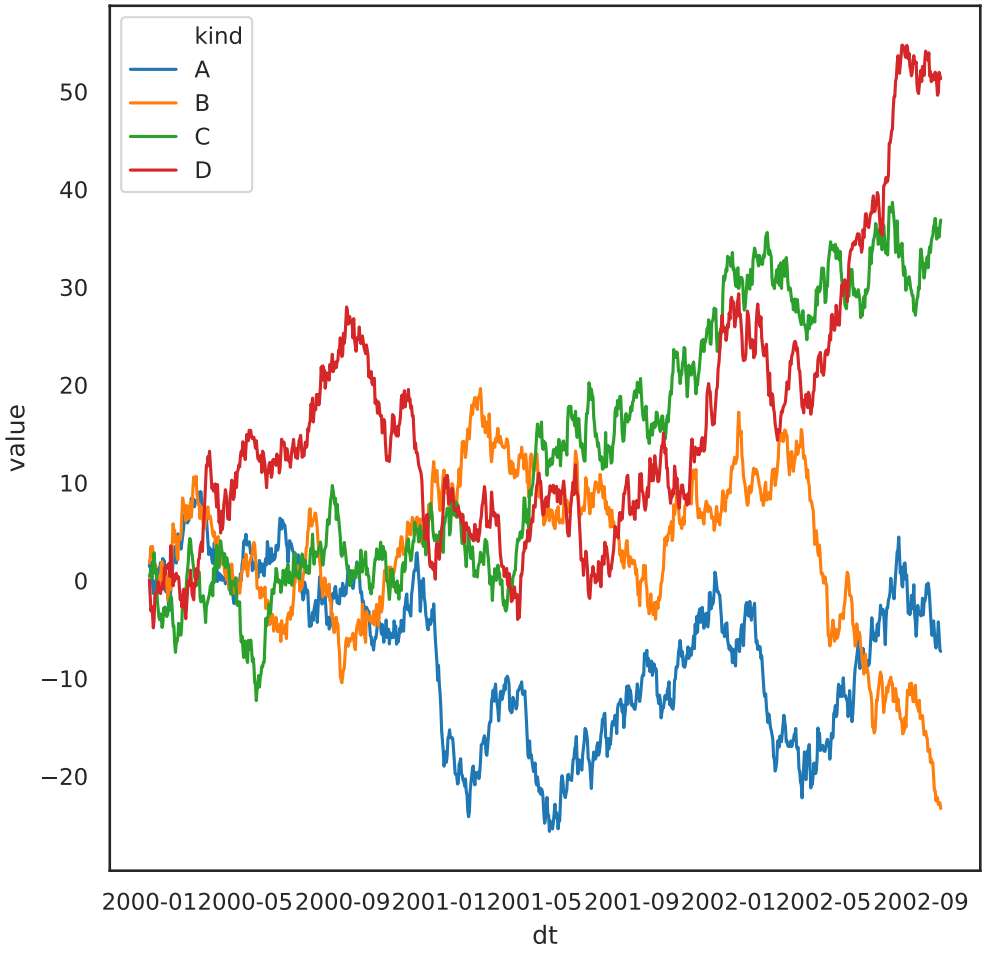
Using `seaborn`...

```
sns.lineplot(data = dfp);
plt.show()
```



However, we can achieve this same plot using the original data, and `seaborn`, in rather short order

```
sns.lineplot(data = ts, x = 'dt', y = 'value', hue = 'kind');
plt.show()
```



In this plot, assigning a variable to hue tells seaborn to draw lines (in this case) of different hues based on values of that variable.

We can use a bit more granular and explicit code for this as well. This allows us a bit more control of the plot.

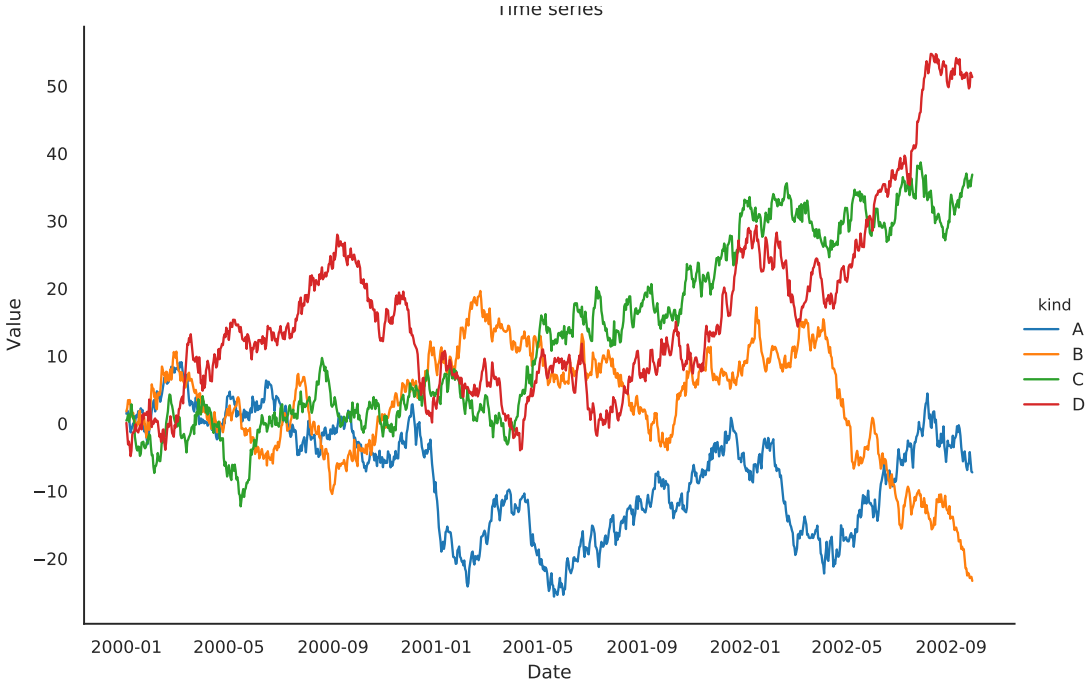
```
g = sns.FacetGrid(ts, hue = 'kind', height = 5, aspect = 1.5)
g.map(plt.plot, 'dt', 'value').add_legend()
```

```
<seaborn.axisgrid.FacetGrid object at 0x137c743a0>
```

```
g.ax.set(xlabel = 'Date',
         ylabel = 'Value',
         title = 'Time series');
```

```
[Text(33.48131111111111, 0.5, 'Value'), Text(0.5, 28.4, 'Date'), Text(0.5, 1.0, 'Time serie
```

```
plt.show()

## All of this code chunk needs to be run at one time, otherwise you get weird errors. T
## is true for many plotting commands which are composed of multiple commands.
```



The `FacetGrid` tells `seaborn` that we're going to layer graphs, with layers based on `hue` and the hues being determined by values of `kind`. Notice that we can add a few more details like the aspect ratio of the plot and so on. The documentation for FacetGrid, which we will also use for facets below, may be helpful in finding all the options you can control.

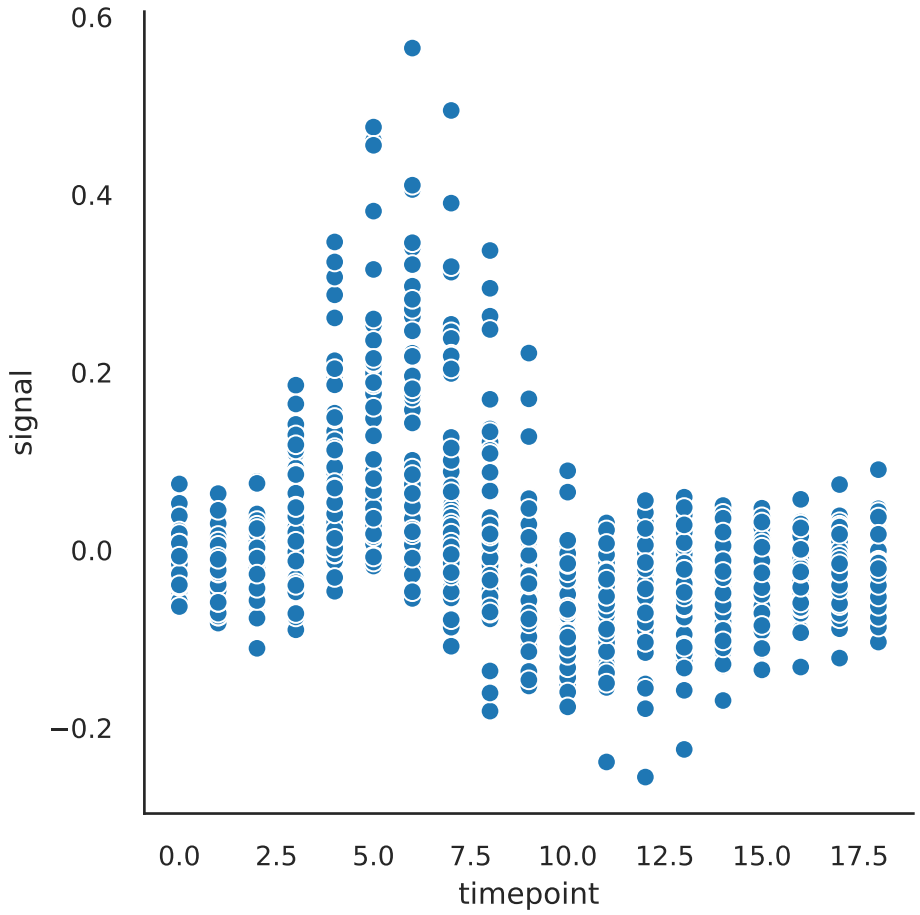
We can also show more than one kind of layer on a single graph

```
fmri = sns.load_dataset('fmri')
```

```
plt.style.use('seaborn-notebook')
sns.relplot(x = 'timepoint', y = 'signal', data = fmri);
```

```
<seaborn.axisgrid.FacetGrid object at 0x1381c3dc0>
```

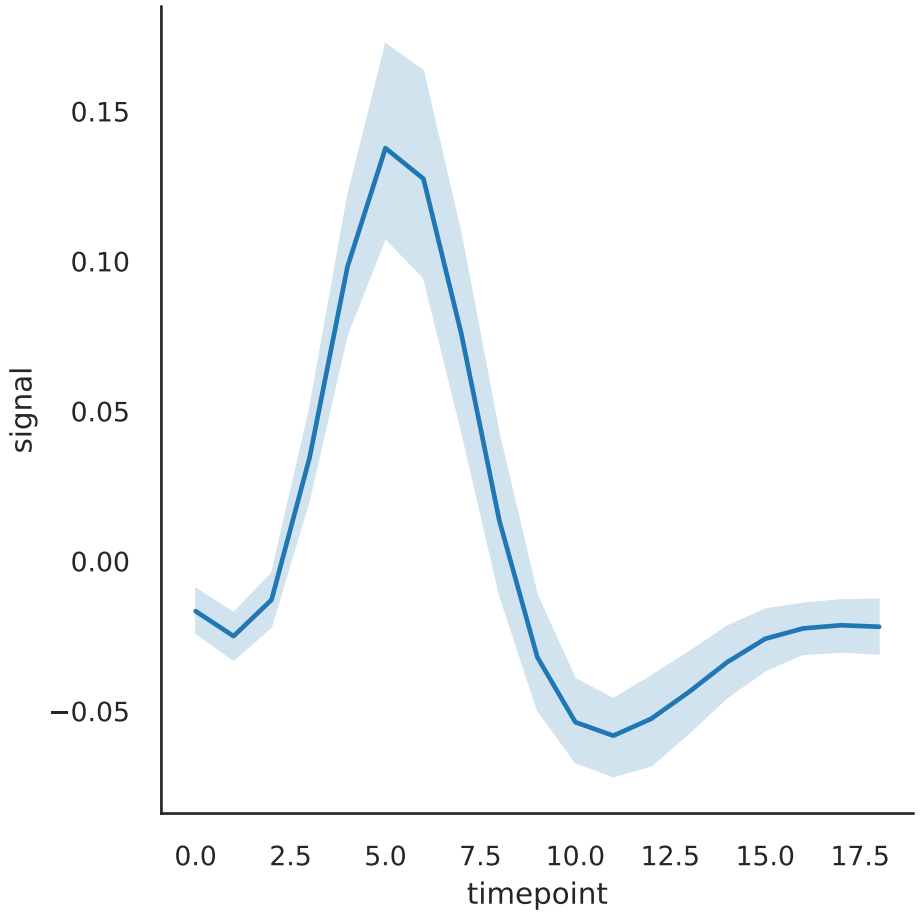
```
plt.show()
```




```
sns.relplot(x = 'timepoint', y = 'signal', data = fmri, kind = 'line');
```

```
<seaborn.axisgrid.FacetGrid object at 0x1384ad040>
```

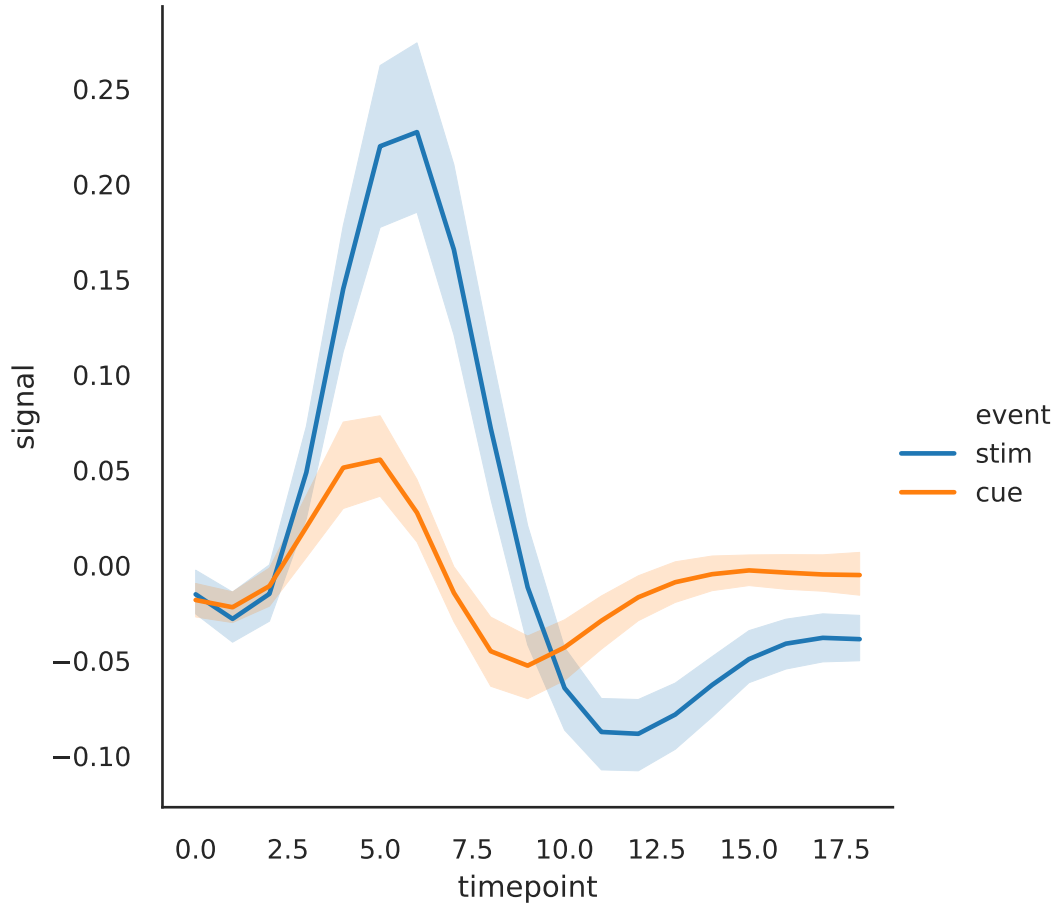
```
plt.show()
```




```
sns.relplot(x = 'timepoint', y = 'signal', data = fmri, kind = 'line', hue ='event');
```

```
<seaborn.axisgrid.FacetGrid object at 0x138fcf820>
```

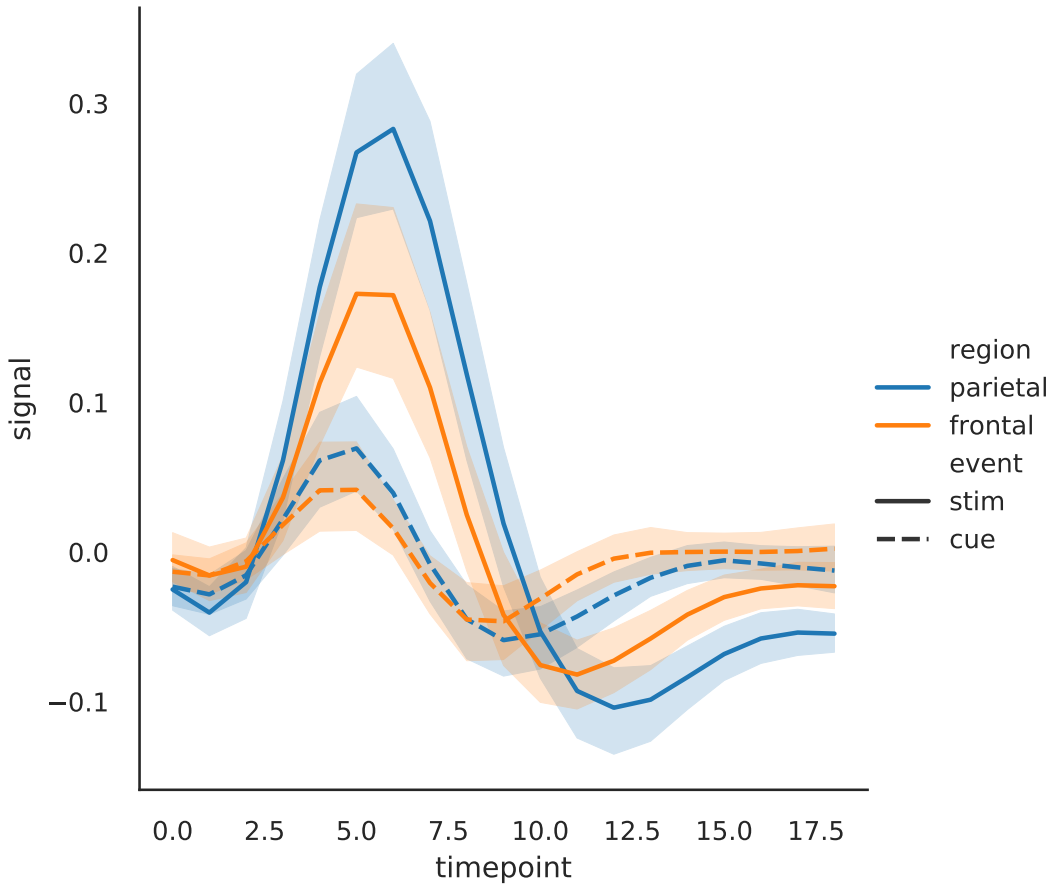
```
plt.show()
```




```
sns.relplot(x = 'timepoint', y = 'signal', data = fmri, hue = 'region',
            style = 'event', kind = 'line');
```

```
<seaborn.axisgrid.FacetGrid object at 0x138441fd0>
```

```
plt.show()
```



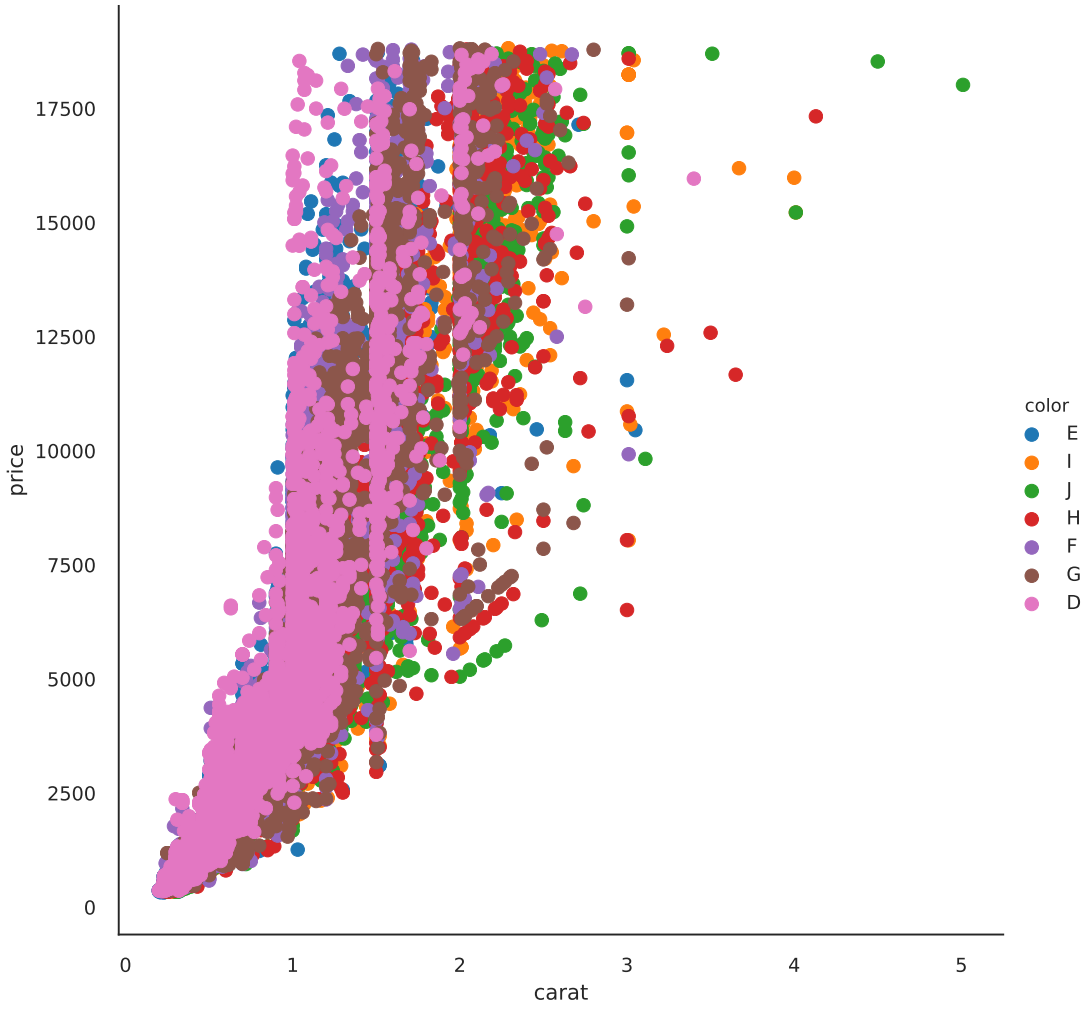
Here we use color to show the region, and line style (solid vs dashed) to show the event.

5.5.0.1 Scatter plots by group

```
g = sns.FacetGrid(diamonds, hue = 'color', height = 7.5)
g.map(plt.scatter, 'carat', 'price').add_legend();
```

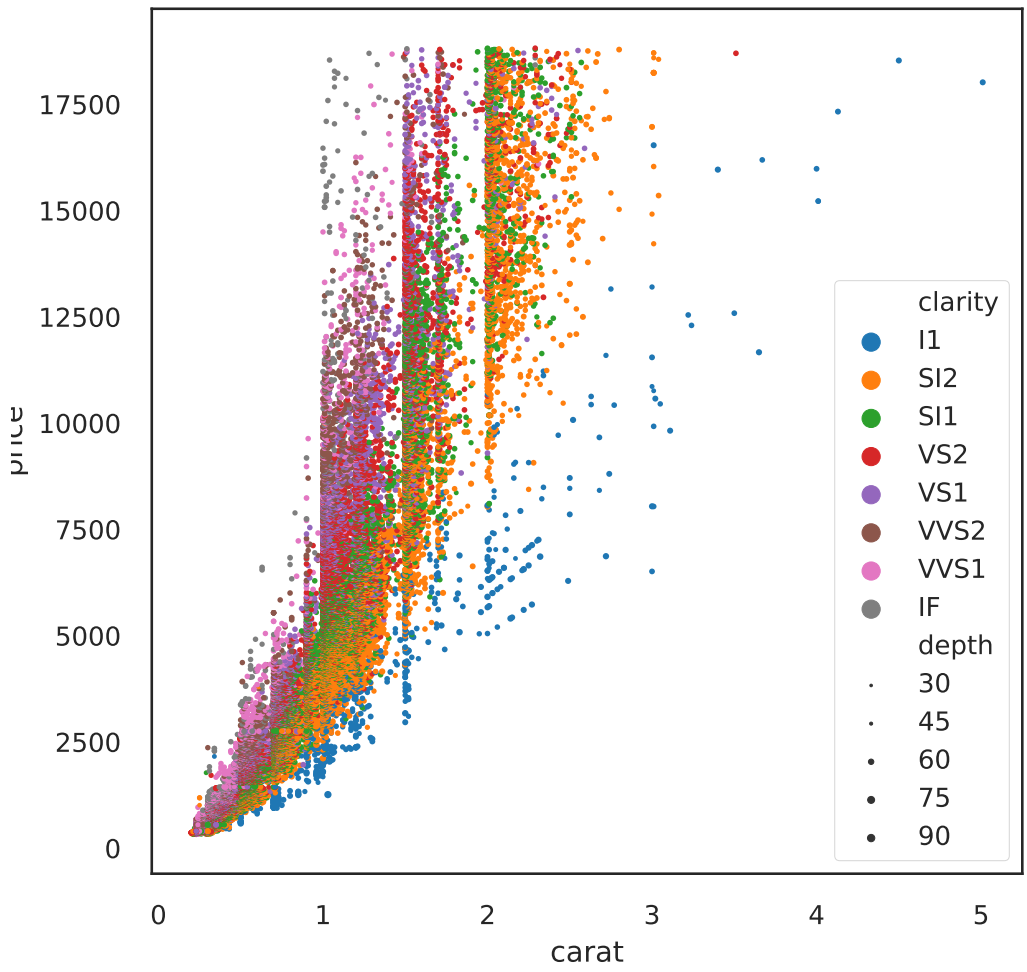
```
<seaborn.axisgrid.FacetGrid object at 0x1390add60>
```

```
plt.show()
```



Notice that this arranges the colors and values for the `color` variable in random order. If we have a preferred order we can impose that using the option `hue_order`.

```
clarity_ranking = ["I1", "SI2", "SI1", "VS2", "VS1", "VVS2", "VVS1", "IF"]
sns.scatterplot(x="carat", y="price",
                hue="clarity", size="depth",
                hue_order=clarity_ranking,
                sizes=(1, 8), linewidth=0,
                data=diamonds);
plt.show()
```



5.5.1 Facets

Facets or trellis graphics is a visualization method where we draw multiple plots in a grid, with each plot corresponding to unique values of a particular variable or combinations of variables. This has also been called *small multiples*.

We'll proceed with an example using the `iris` dataset.

```
iris = pd.read_csv('data/iris.csv')
iris.head()
```

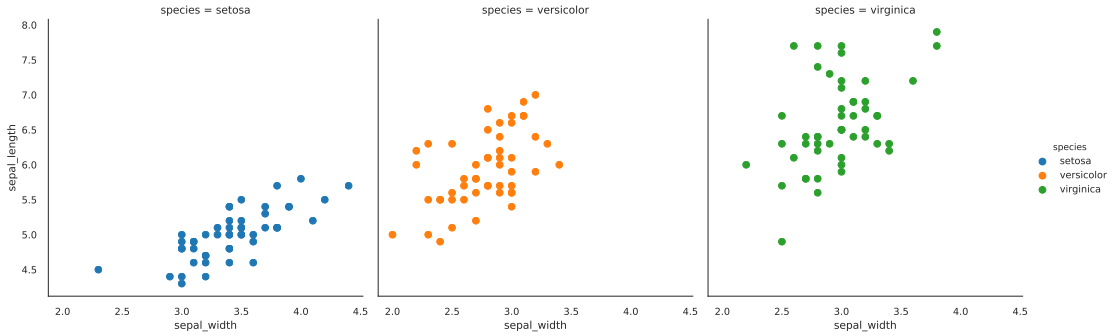
```
   sepal_length  sepal_width  petal_length  petal_width species
0           5.1          3.5           1.4          0.2  setosa
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
```

```
3            4.6           3.1           1.5           0.2  setosa
4            5.0           3.6           1.4           0.2  setosa
```

```
g = sns.FacetGrid(iris, col = 'species', hue = 'species', height = 5)
g.map(plt.scatter, 'sepal_width', 'sepal_length').add_legend();
```

```
<seaborn.axisgrid.FacetGrid object at 0x112662370>
```

```
plt.show()
```



Here we use `FacetGrid` to indicate that we're creating multiple subplots by specifying the option `col` (for column). So this code says we are going to create one plot per level of species, arranged as separate columns (or in effect along one row). You could also specify `row` which would arrange the plots one to a row, or, in effect, in one column.

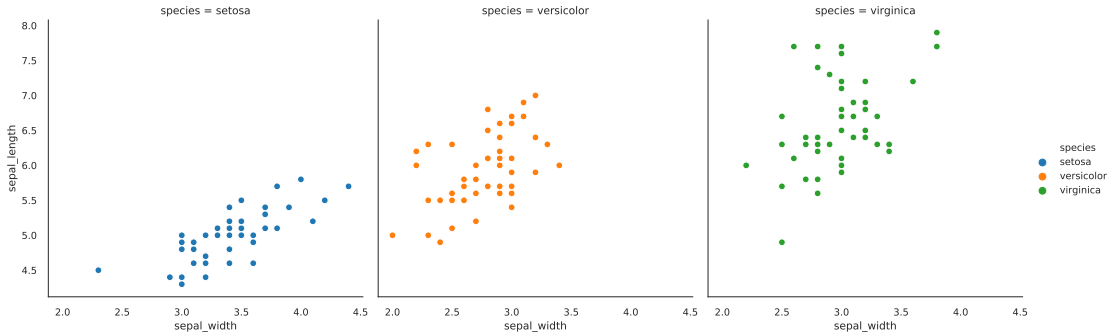
The `map` function says, take the facets I've defined and stored in `g`, and in each one, plot a scatter plot with `sepal_width` on the x-axis and `sepal_length` on the y-axis.

We could also use `relplot` for a more compact solution.

```
sns.relplot(x = 'sepal_width', y = 'sepal_length', data = iris,
            col = 'species', hue = 'species');
```

```
<seaborn.axisgrid.FacetGrid object at 0x138a584c0>
```

```
plt.show()
```

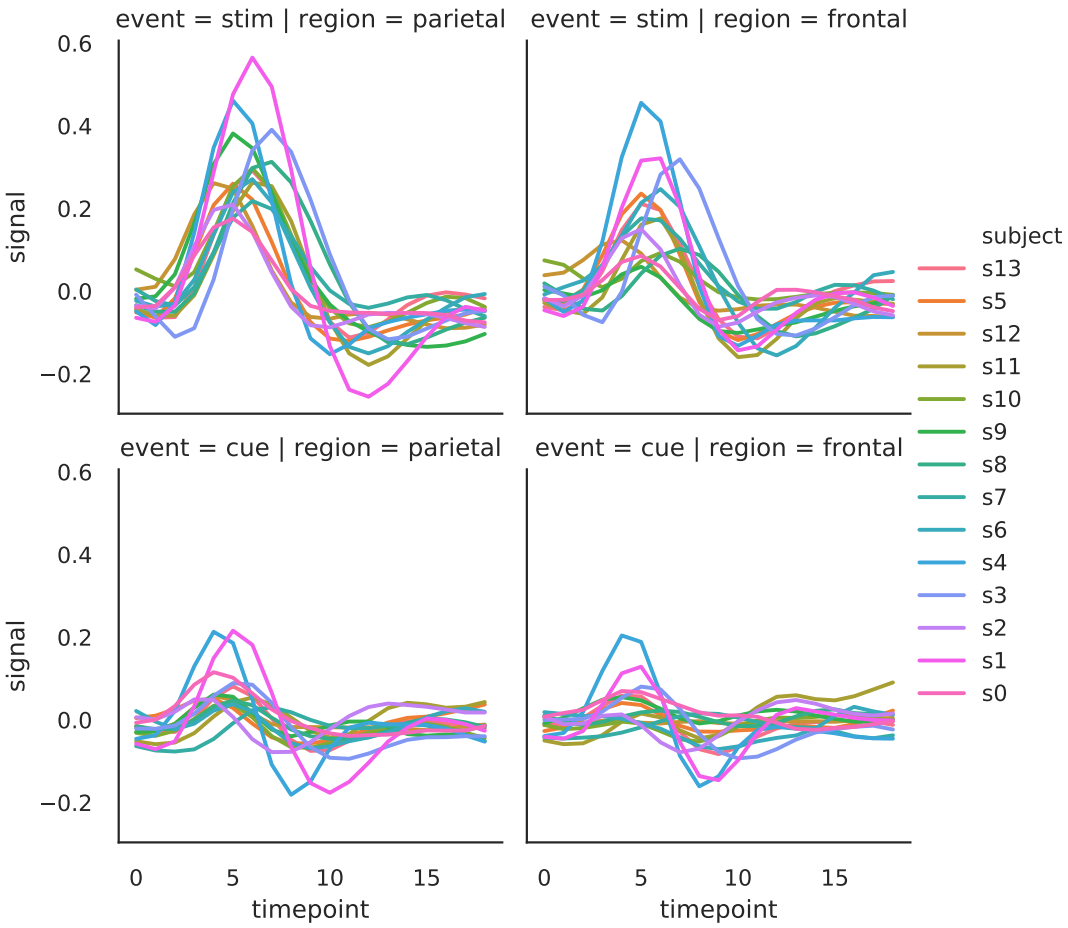


A bit more of a complicated example, using the `fmri` data, where we're coloring lines based on the subject, and creating a 2-d grid, where region of the brain in along columns and event type is along rows.

```
sns.relplot(x="timepoint", y="signal", hue="subject",
            col="region", row="event", height=3,
            kind="line", estimator=None, data=fmri);
```

```
<seaborn.axisgrid.FacetGrid object at 0x127c2de80>
```

```
plt.show()
```

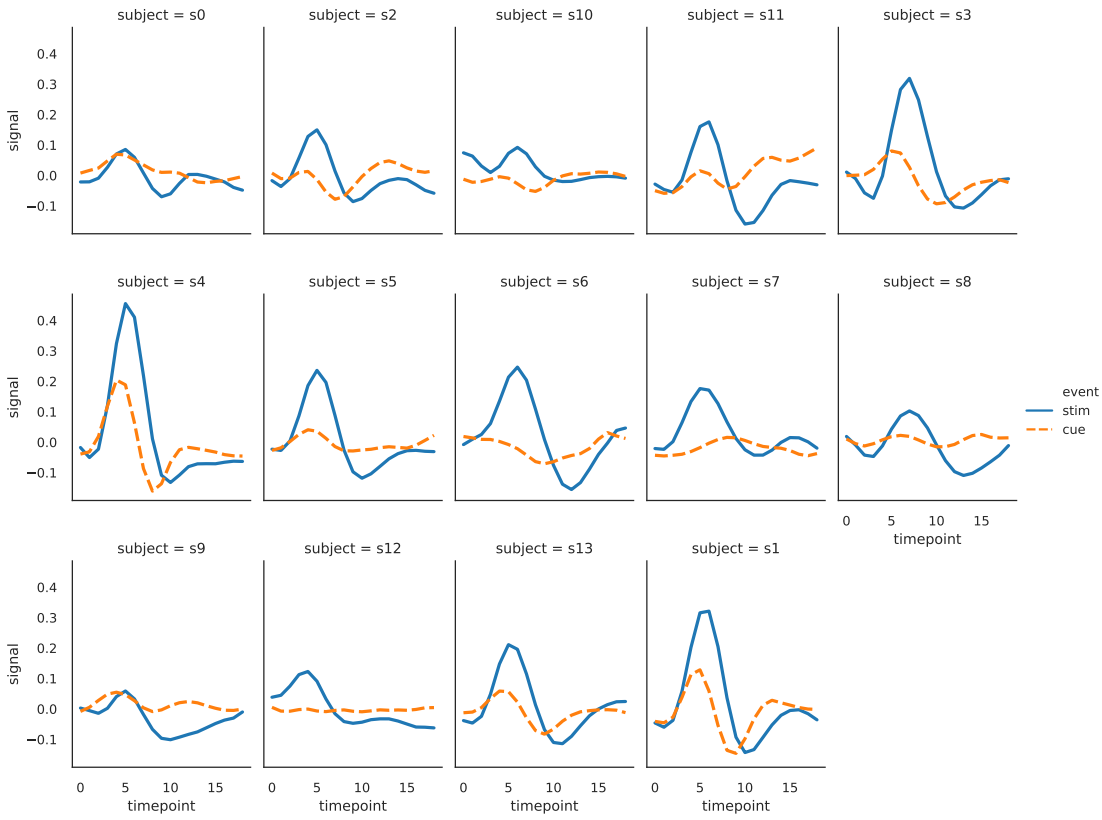


In the following example, we want to show how each subject fares for each of the two events, just within the frontal region. We let `seaborn` figure out the layout, only specifying that we'll be going along rows ("by column") and also saying we'll wrap around to the beginning once we've got to 5 columns. Note we use the `query` function to filter the dataset.

```
sns.relplot(x="timepoint", y="signal", hue="event", style="event",
            col="subject", col_wrap=5,
            height=3, aspect=.75, linewidth=2.5,
            kind="line", data=fmri.query("region == 'frontal'"));
```

```
<seaborn.axisgrid.FacetGrid object at 0x13965b8e0>
```

```
plt.show()
```

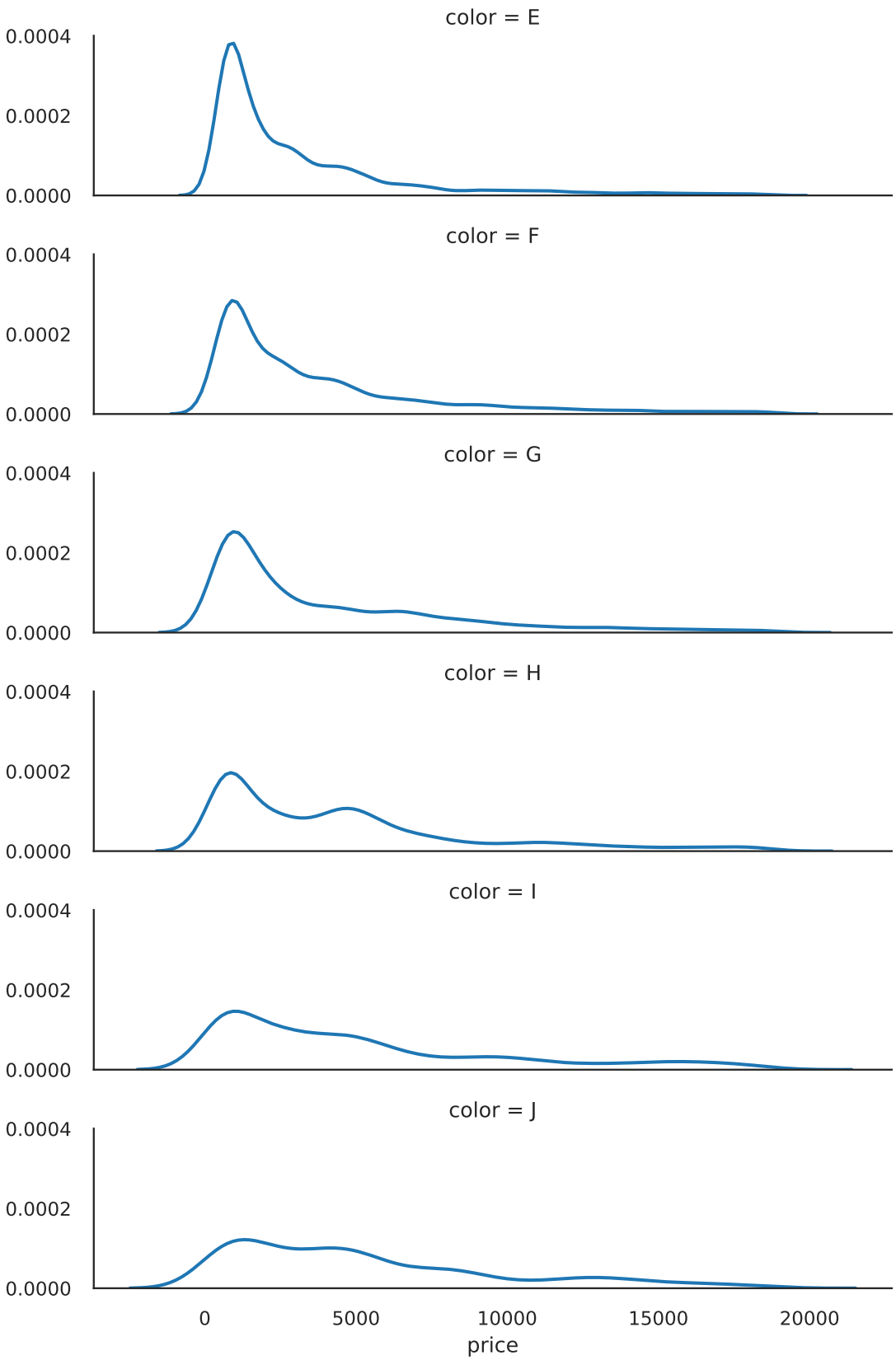


In the following example we want to compare the distribution of price from the diamonds dataset by color, and so it makes sense to create density plots of the price distribution and stack them one below the next so we can visually compare them.

```
ordered_colors = ['E','F','G','H','I','J']
g = sns.FacetGrid(data = diamonds, row = 'color', height = 1.7,
                  aspect = 4, row_order = ordered_colors)
g.map(sns.kdeplot, 'price');
```

```
<seaborn.axisgrid.FacetGrid object at 0x1381c4ee0>
```

```
plt.show()
```


color = E
color = F
color = G
color = H
color = I
color = J
0.0004
0.0002
0.0000
0
5000
10000
15000
20000
price

You need to use `FacetGrid` to create sets of univariate plots since there is no particular method that allows univariate plots over a grid like `relplot` for bivariate plots.

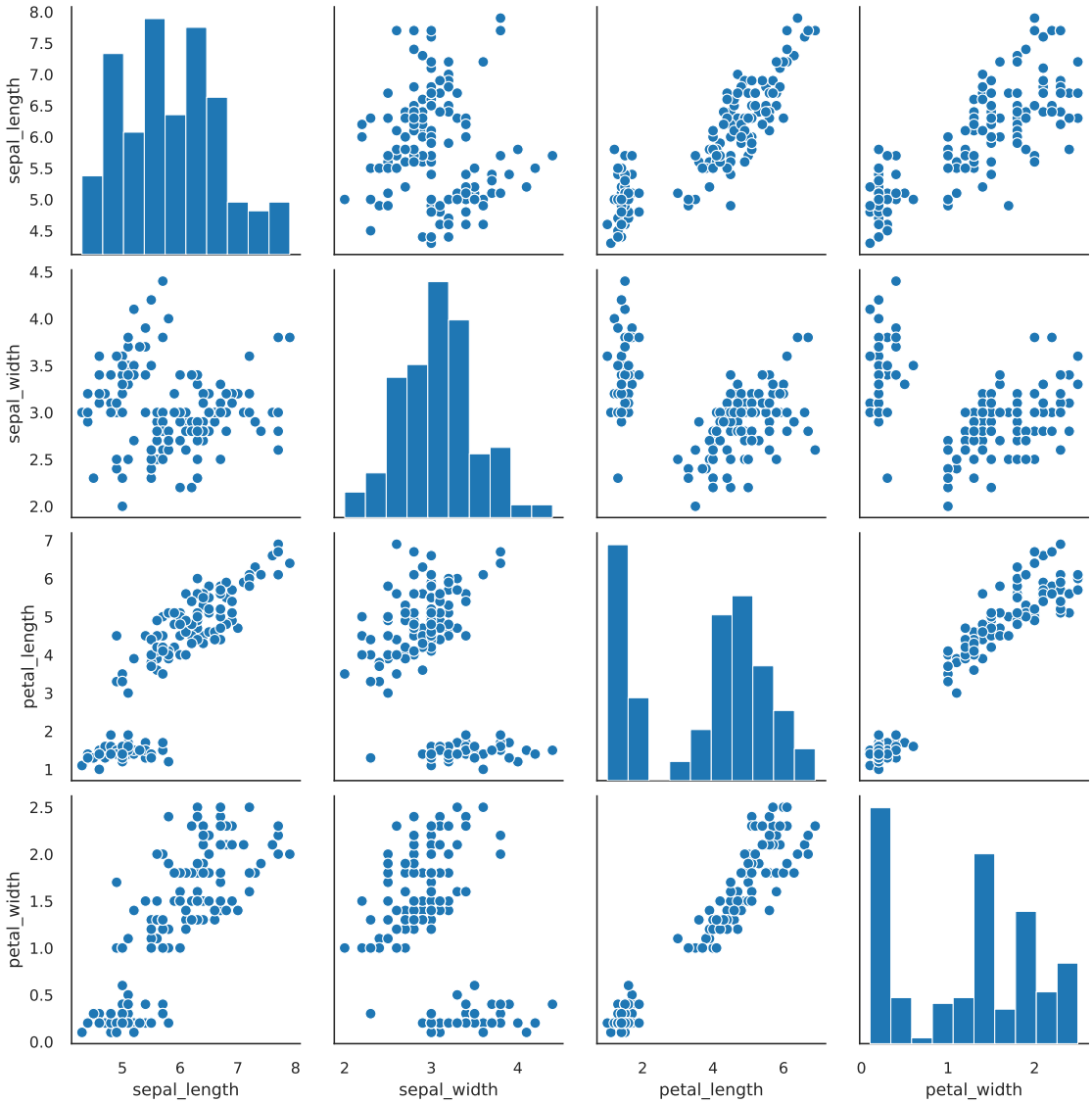
5.5.2 Pairs plots

The pairs plot is a quick way to compare every pair of variables in a dataset (or at least, every pair of continuous variables) in a grid. You can specify what kind of univariate plot will be displayed on the diagonal locations on the grid, and which bivariate plots will be displayed on the off-diagonal locations.

```
sns.pairplot(data=iris);
```

```
<seaborn.axisgrid.PairGrid object at 0x127e543a0>
```

```
plt.show()
```



You can achieve more customization using `PairGrid`.

```
g = sns.PairGrid(iris, diag_sharey=False);
g.map_upper(sns.scatterplot);
```

```
<seaborn.axisgrid.PairGrid object at 0x127c9fac0>
```

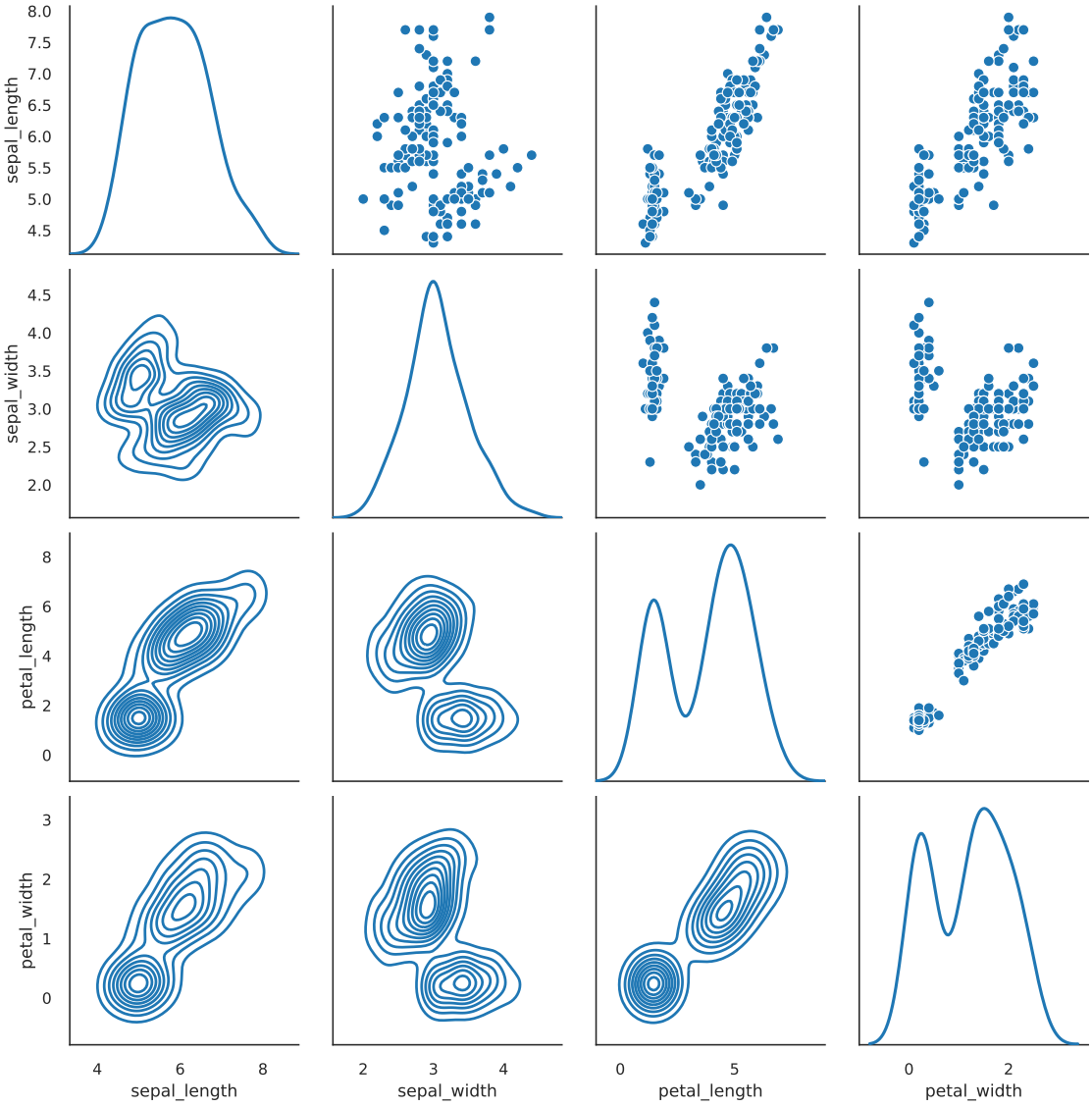
```
g.map_lower(sns.kdeplot, colors="C0");
```

```
<seaborn.axisgrid.PairGrid object at 0x127c9fac0>
```

```
g.map_diag(sns.kdeplot, lw=2);
```

```
<seaborn.axisgrid.PairGrid object at 0x127c9fac0>
```

```
plt.show()
```



5.6 Customizing the look

5.6.1 Themes

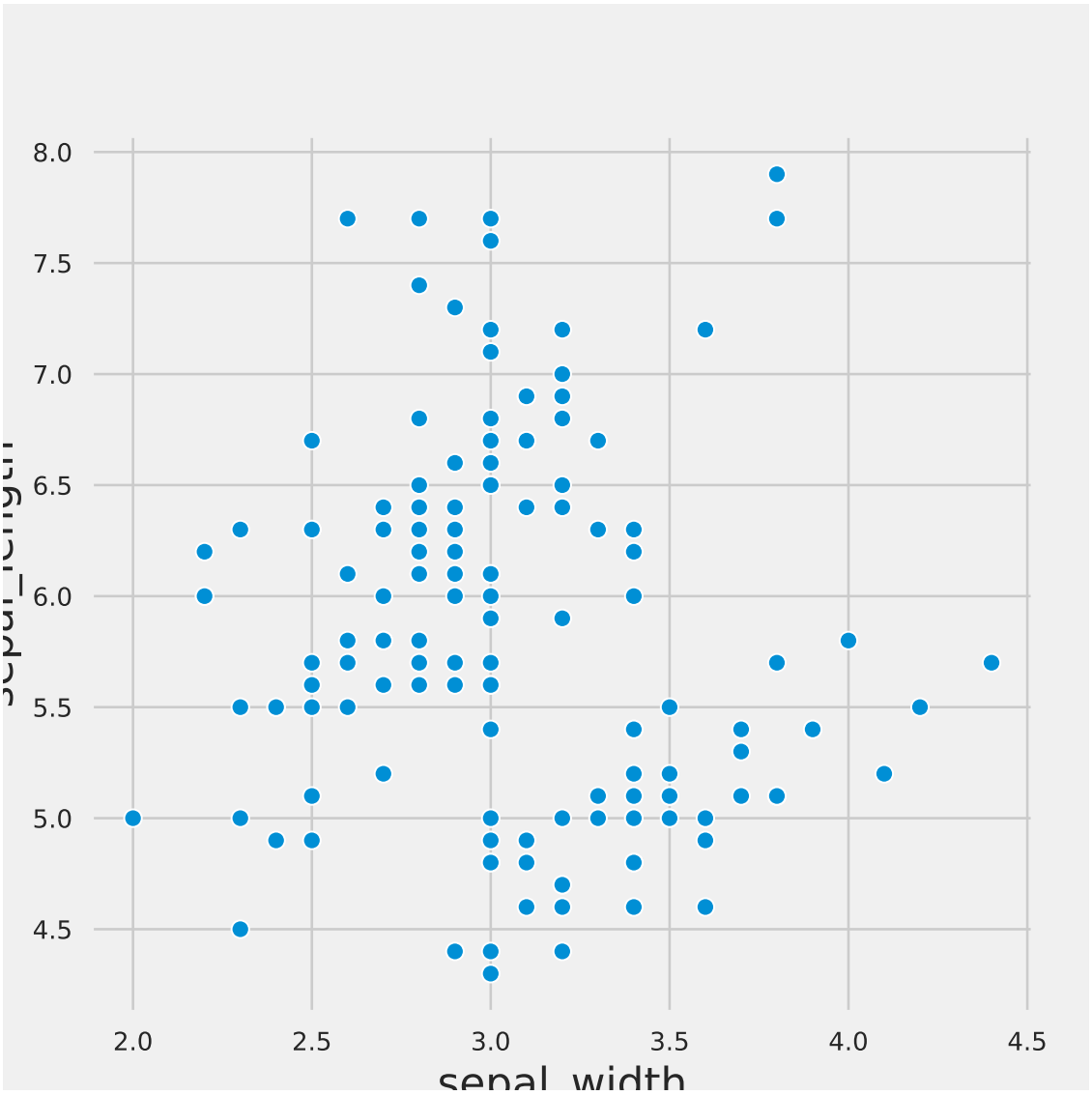
There are several themes available in the modern `matplotlib`, some of which borrow from `seaborn`. You can see the available themes and play around.

```
plt.style.available
```

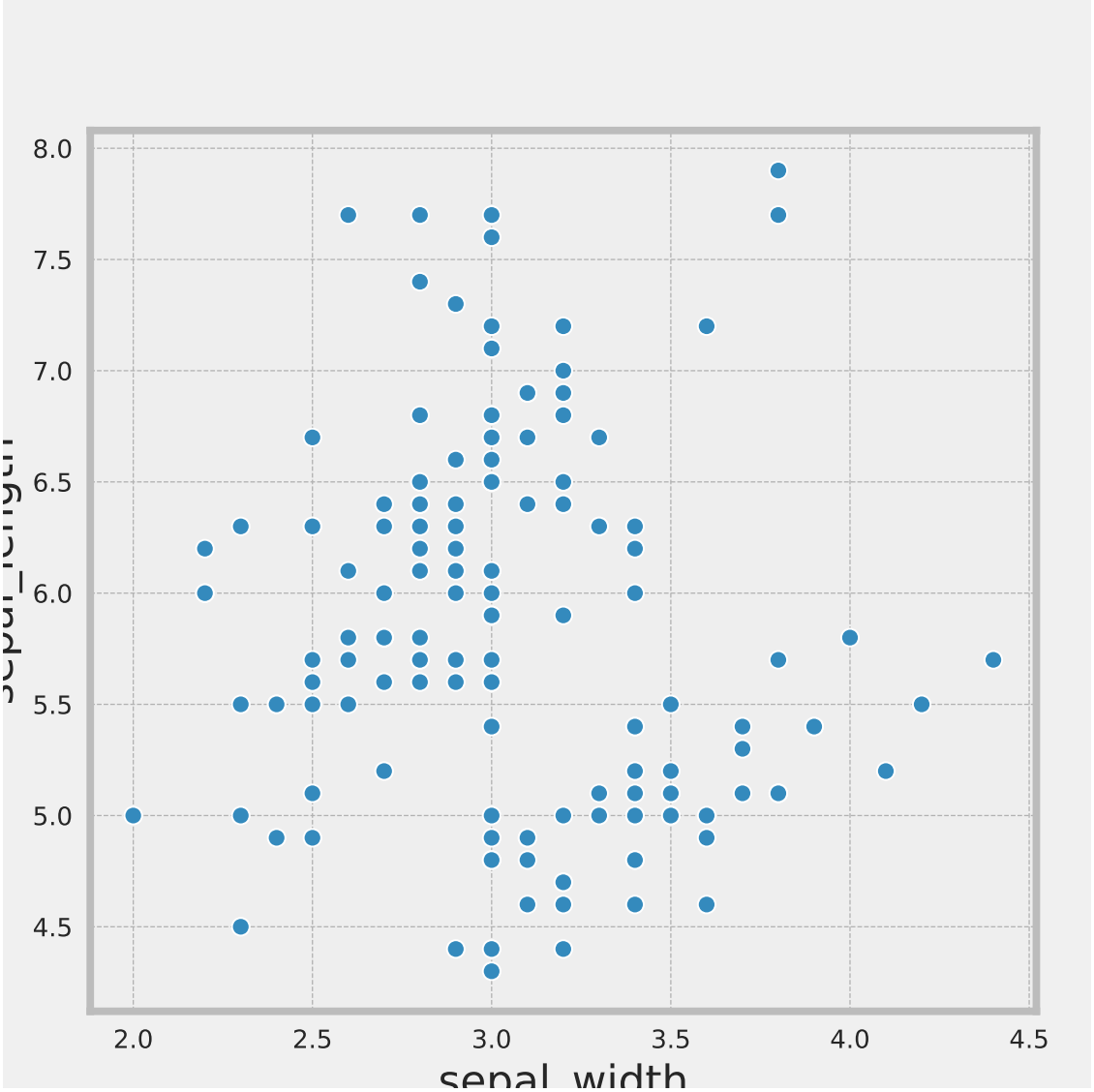
```
['Solarize_Light2', '_classic_test_patch', 'bmh', 'classic', 'dark_background', 'fast', 'f
bright', 'seaborn-colorblind', 'seaborn-dark', 'seaborn-dark-palette', 'seaborn-
darkgrid', 'seaborn-deep', 'seaborn-muted', 'seaborn-notebook', 'seaborn-
paper', 'seaborn-pastel', 'seaborn-poster', 'seaborn-talk', 'seaborn-
ticks', 'seaborn-white', 'seaborn-whitegrid', 'tableau-colorblind10']
```

See some examples below.

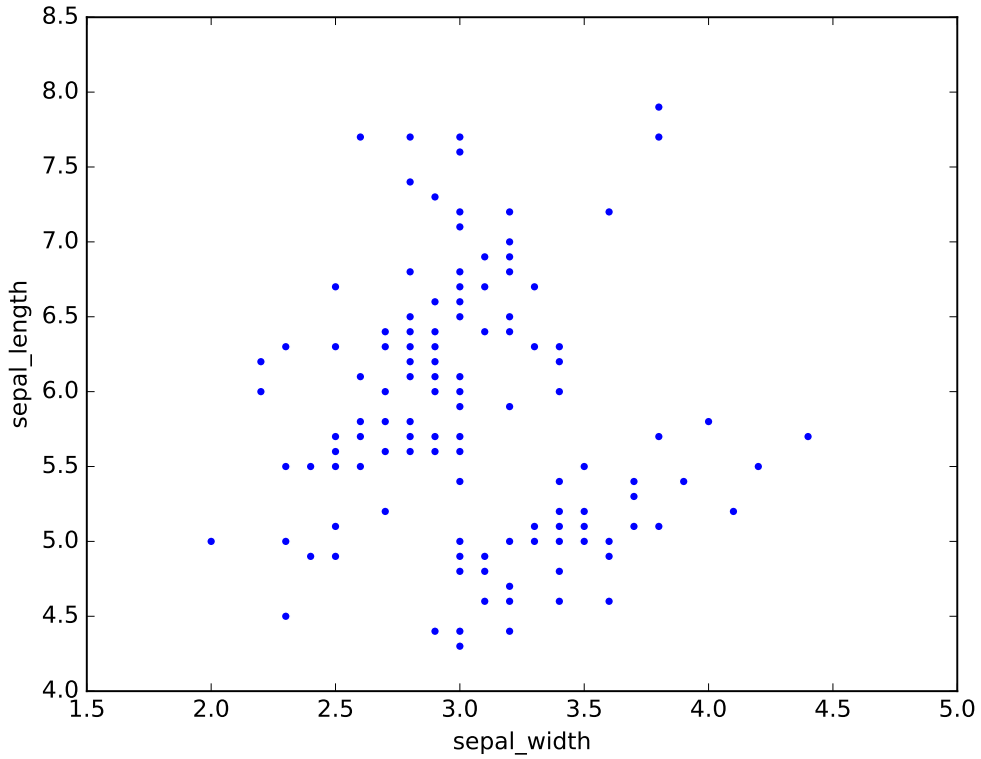
```
plt.style.use('fivethirtyeight')
sns.scatterplot(data = iris, x = 'sepal_width', y = 'sepal_length');
plt.show()
```



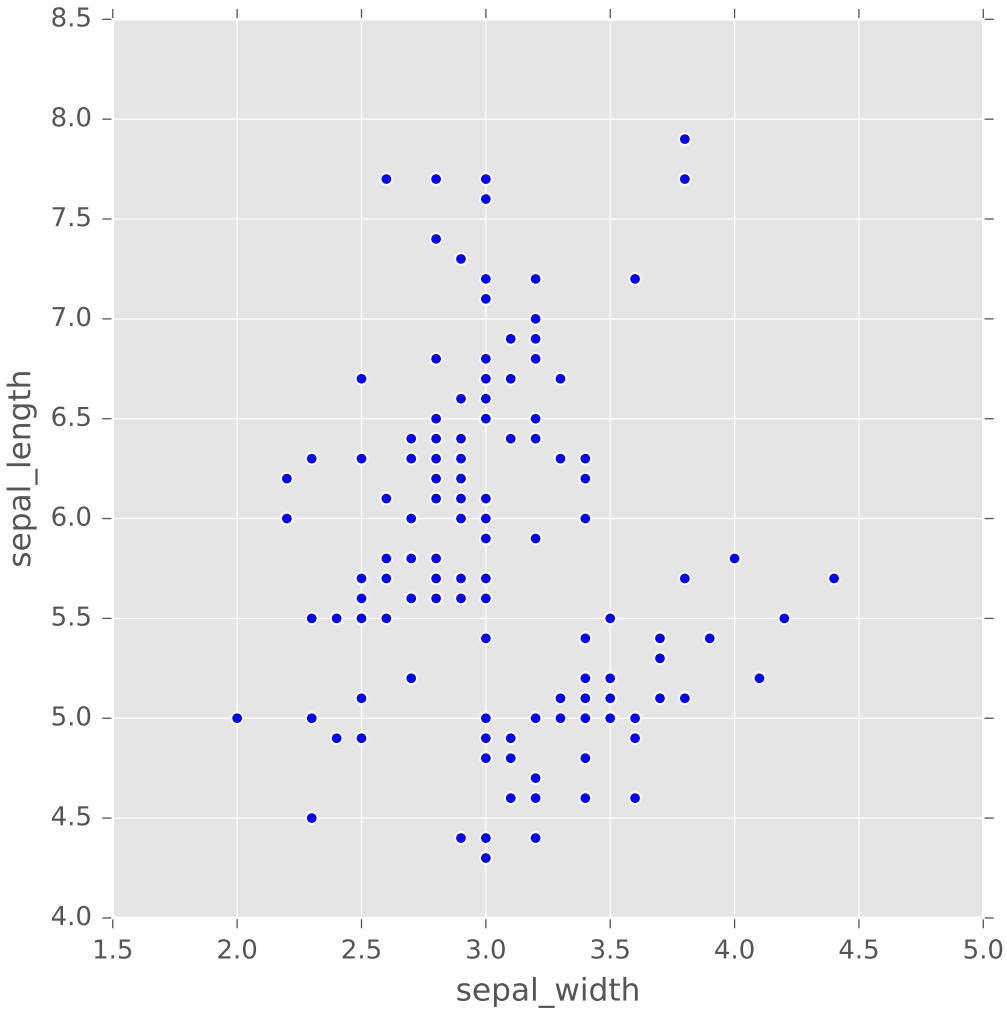

```
plt.style.use('bmh')
sns.scatterplot(data = iris, x = 'sepal_width', y = 'sepal_length');
plt.show()
```



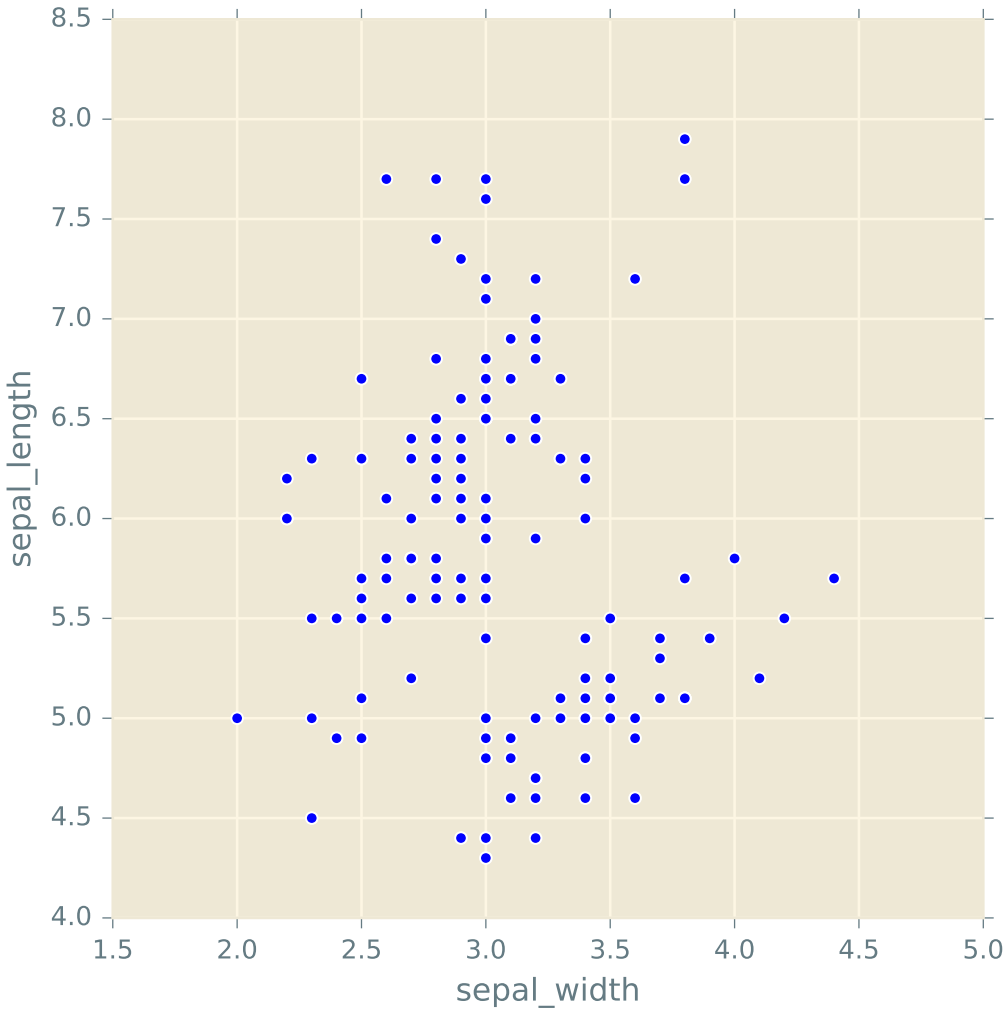

```
plt.style.use('classic')
sns.scatterplot(data = iris, x = 'sepal_width', y = 'sepal_length');
plt.show()
```




```
plt.style.use('ggplot')
sns.scatterplot(data = iris, x = 'sepal_width', y = 'sepal_length');
plt.show()
```




```
plt.style.use('Solarize_Light2')
sns.scatterplot(data = iris, x = 'sepal_width', y = 'sepal_length');
plt.show()
```



One small syntax point. You may have noticed in your own work that you get a little annoying line in the output when you plot. You can prevent that from happening by putting a semi-colon (`;`) after the last plotting command

5.7 Finer control with matplotlib

As you can see from the figure, you can control each aspect of the plot displayed above using `matplotlib`. I won't go into the details, and will leave it to you to look at the `matplotlib` documentation and examples if you need to customize at this level of granularity.

The following is an example using pure `matplotlib`. You can see how you can build up a plot. The crucial part here is that you need to run the code from each chunk together.

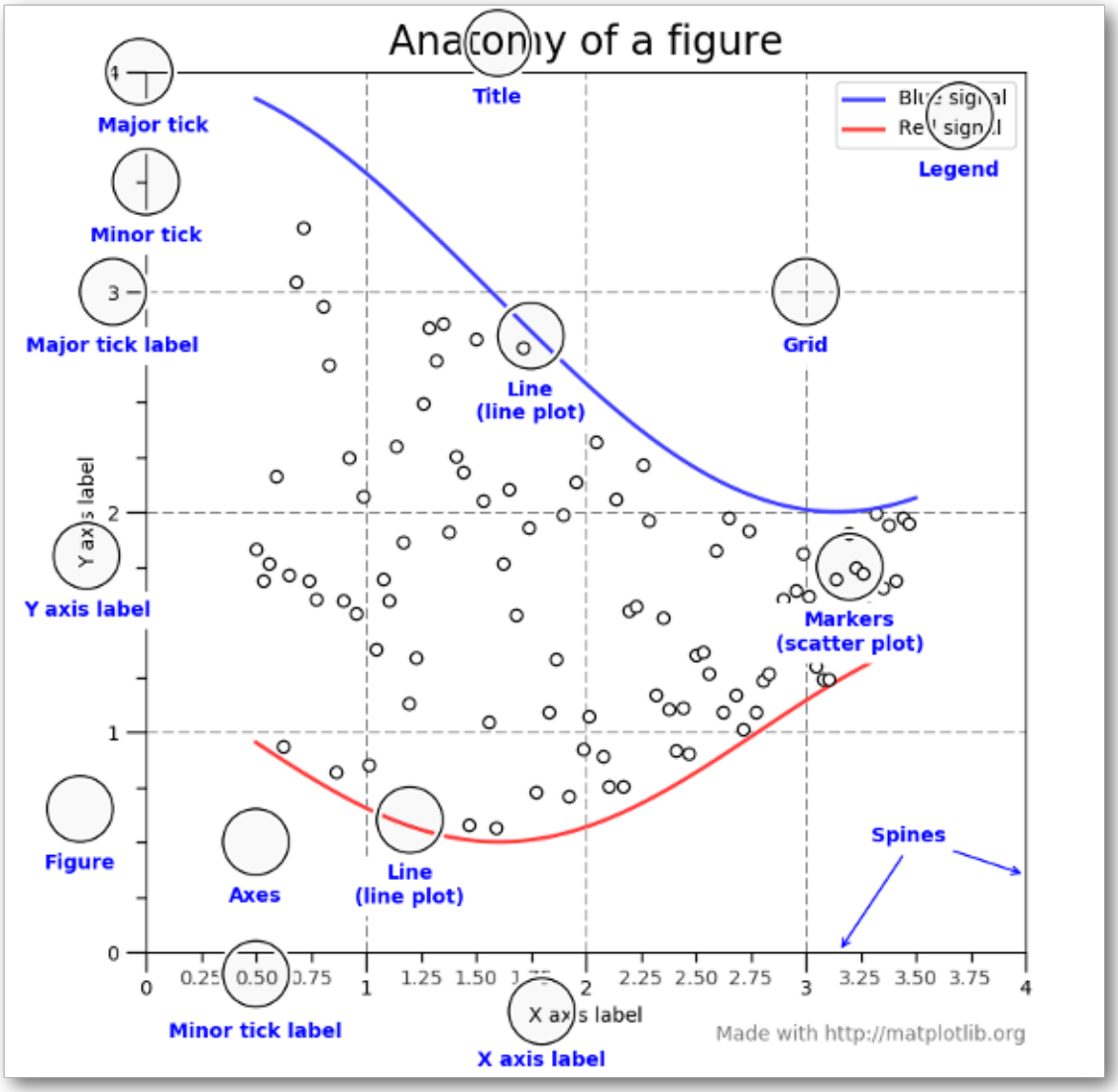
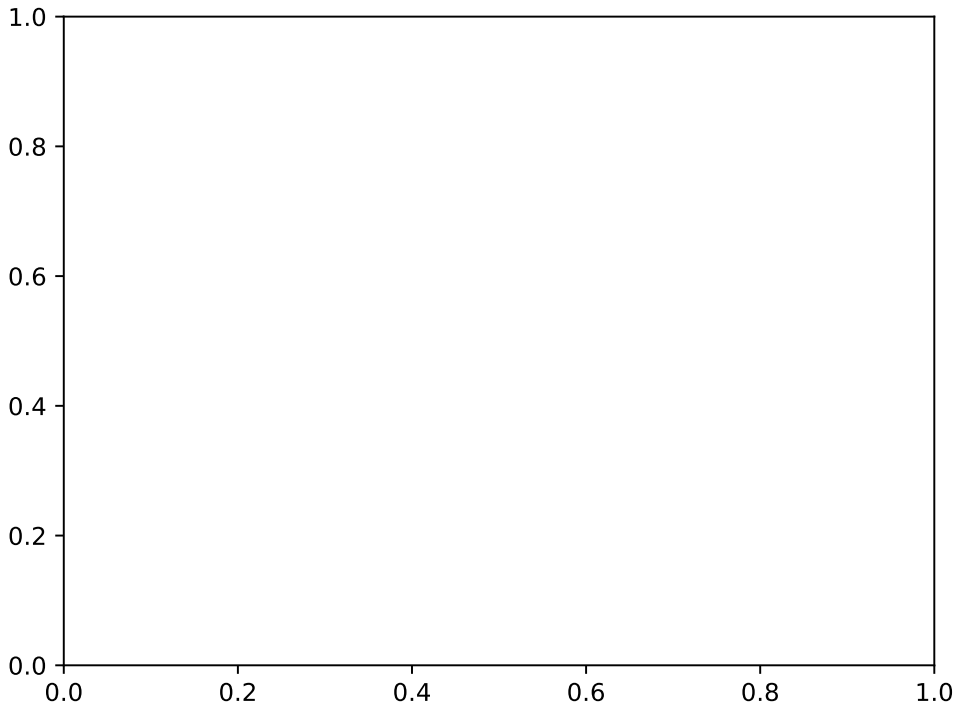


Figure 5.1: https://matplotlib.org/tutorials/introductory/usage.html#sphx-glr-tutorials-introductory-usage-py

```
from matplotlib.ticker import FuncFormatter

data = {'Barton LLC': 109438.50,
        'Frami, Hills and Schmidt': 103569.59,
        'Fritsch, Russel and Anderson': 112214.71,
        'Jerde-Hilpert': 112591.43,
        'Keeling LLC': 100934.30,
        'Koepp Ltd': 103660.54,
        'Kulas Inc': 137351.96,
        'Trantow-Barrows': 123381.38,
        'White-Trantow': 135841.99,
        'Will LLC': 104437.60}
group_data = list(data.values())
group_names = list(data.keys())
group_mean = np.mean(group_data)
```

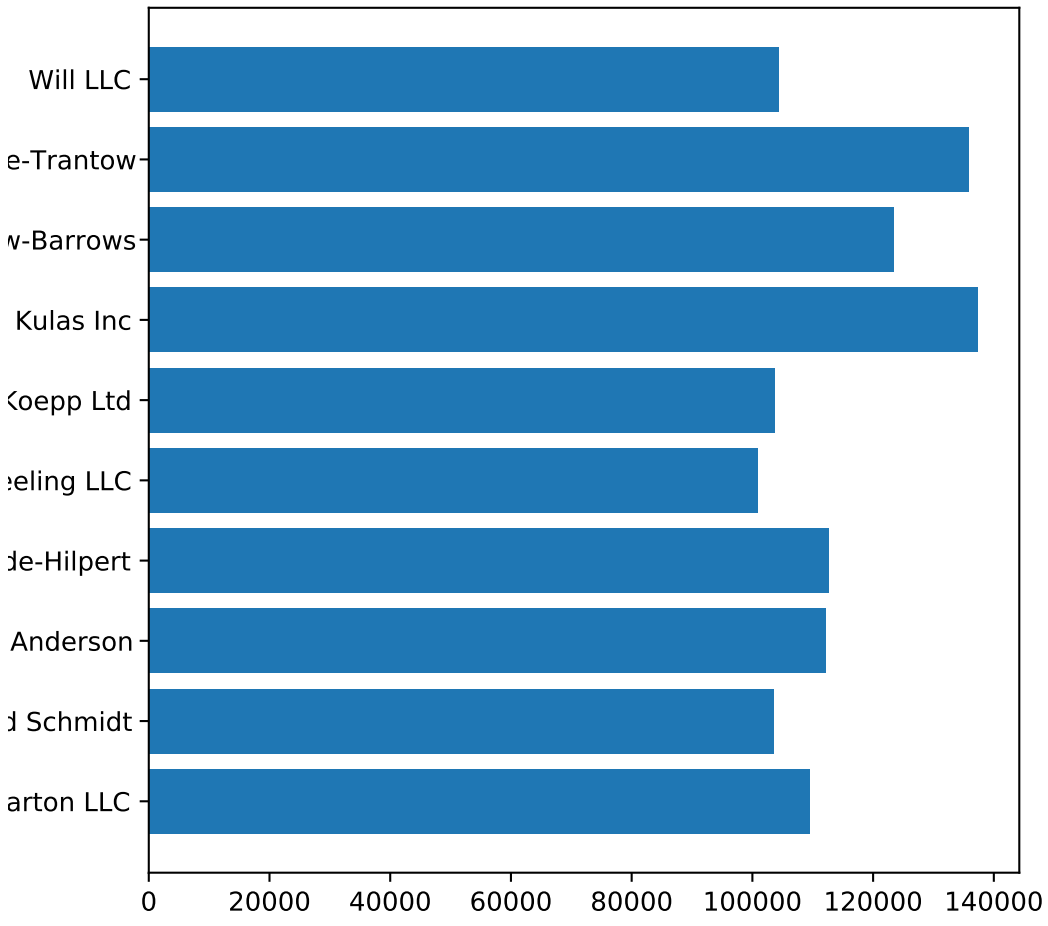
```
plt.style.use('default')
fig, ax = plt.subplots()
plt.show()
```



```
fig, ax = plt.subplots()
ax.barh(group_names, group_data);
```

```
<BarContainer object of 10 artists>
```

```
plt.show()
```



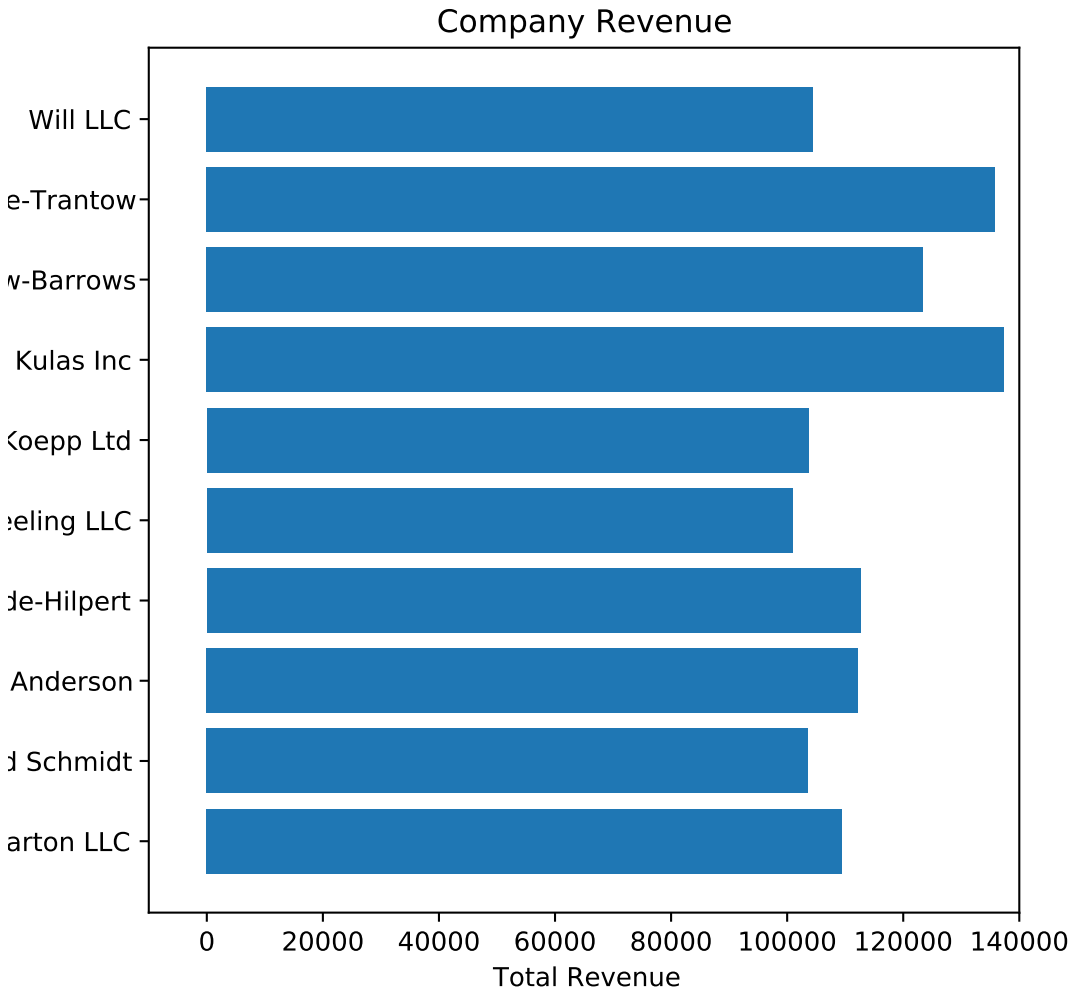

```
fig, ax = plt.subplots()
ax.barh(group_names, group_data)
```

```
<BarContainer object of 10 artists>
```

```
ax.set(xlim = [-10000, 140000], xlabel = 'Total Revenue', ylabel = 'Company',
       title = 'Company Revenue');
```

```
[Text(0, 0.5, 'Company'), (-10000.0, 140000.0), Text(0.5, 0, 'Total Revenue'), Text(0.5, 1.0
```

```
plt.show()
```




```
fig, ax = plt.subplots(figsize=(8, 4))
ax.barh(group_names, group_data)
```

```
<BarContainer object of 10 artists>
```

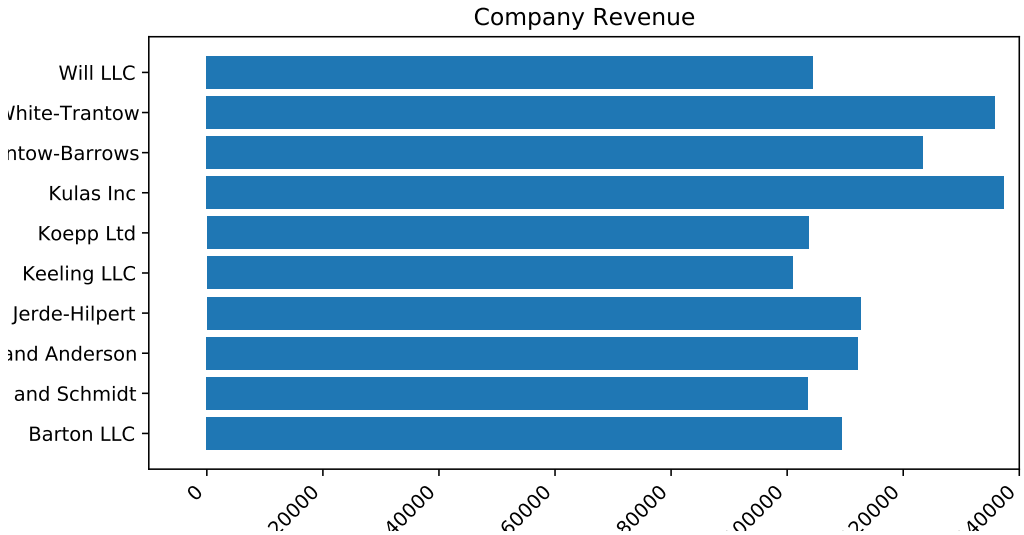
```
labels = ax.get_xticklabels()
plt.setp(labels, rotation=45, horizontalalignment='right')
```

```
[None, None, None, None, None, None, None, None, None, None, None, None, None, None, No
```

```
ax.set(xlim=[-10000, 140000], xlabel='Total Revenue', ylabel='Company',
       title='Company Revenue');
```

```
[Text(0, 0.5, 'Company'), (-10000.0, 140000.0), Text(0.5, 0, 'Total Revenue'), Text(0.5, 1.0
```

```
plt.show()
```



After you have created your figure, you do need to save it to disk so that you can use it in your Word or Markdown or PowerPoint document. You can see the formats available.

```
fig.canvas.get_supported_filetypes()
```

```
{'ps': 'Postscript', 'eps': 'Encapsulated Postscript', 'pdf': 'Portable Document Format', '
```

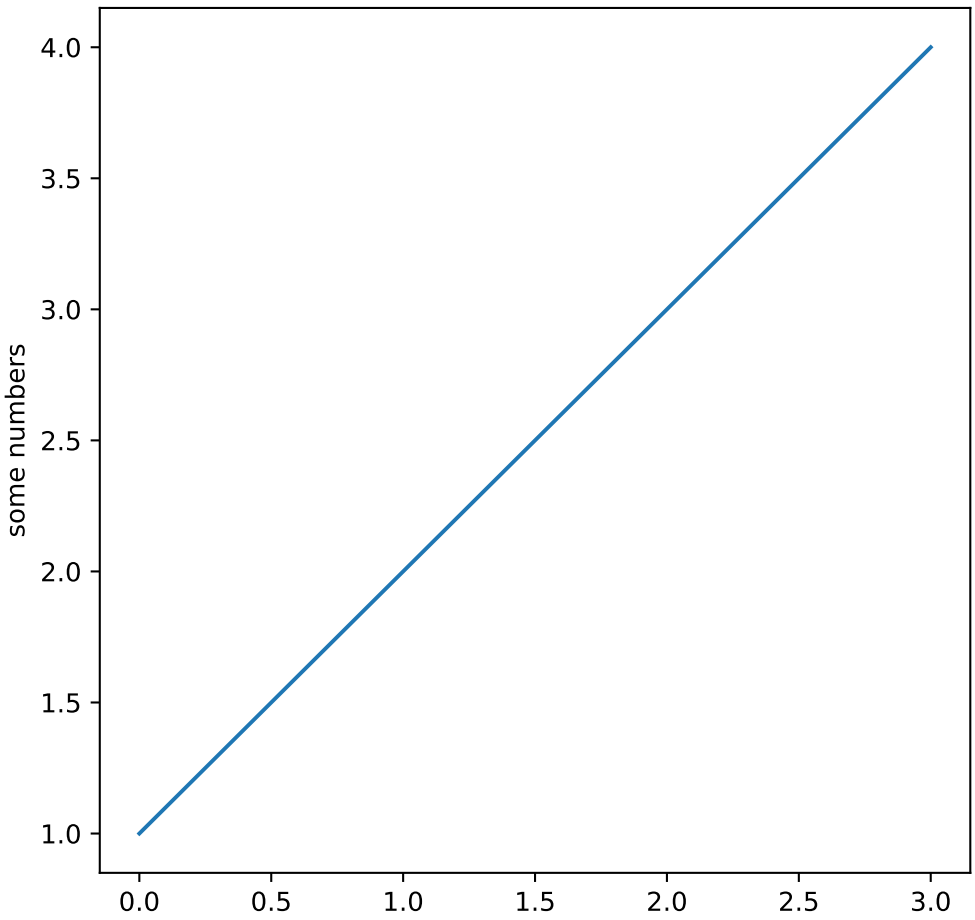
The type will be determined by the ending of the file name. You can add some options depending on the type. I'm showing an example of saving the figure to a PNG file. Typically I'll save figures to a vector graphics format like PDF, and then convert into other formats, since that results in minimal resolution loss. You of course have the option to save to your favorite format.

```
# fig.savefig('sales.png', dpi = 300, bbox_inches = 'tight')
```

5.7.1 Matlab-like plotting

`matplotlib` was originally developed to emulate Matlab. Though this kind of syntax is no longer recommended, it is still available and may be of use to those coming to Python from Matlab or Octave.

```
import matplotlib.pyplot as plt
plt.plot([1, 2, 3, 4]);
plt.ylabel('some numbers');
plt.show()
```

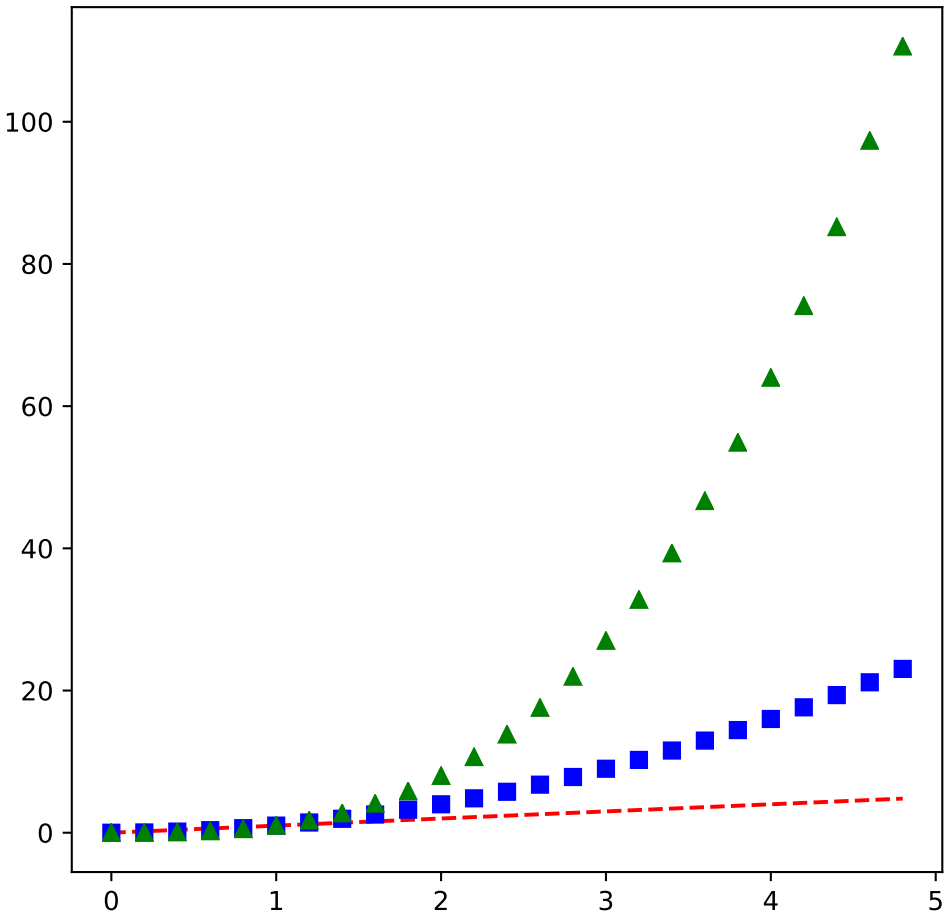



```
import numpy as np

# evenly sampled time at 200ms intervals
t = np.arange(0., 5., 0.2)
```

```
# red dashes, blue squares and green triangles
plt.plot(t, t, 'r--', t, t**2, 'bs', t, t**3, 'g^');
```

```
[<matplotlib.lines.Line2D object at 0x13896fd30>, <matplotlib.lines.Line2D object at 0x1389
```

```
plt.show()
```

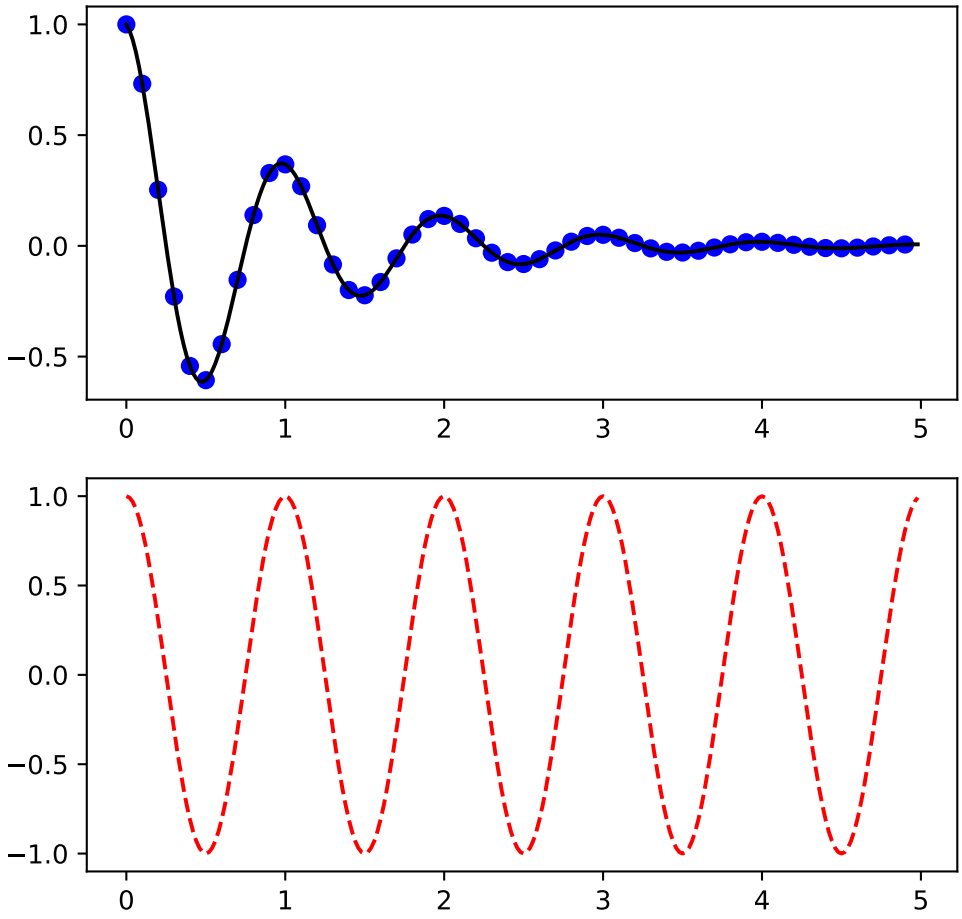



```
def f(t):
    return np.exp(-t) * np.cos(2*np.pi*t)

t1 = np.arange(0.0, 5.0, 0.1)
t2 = np.arange(0.0, 5.0, 0.02)
```

```
plt.figure();
plt.subplot(211);
plt.plot(t1, f(t1), 'bo', t2, f(t2), 'k');
```

```
[<matplotlib.lines.Line2D object at 0x127cf86a0>, <matplotlib.lines.Line2D object at 0x127
```

```
plt.subplot(212);
plt.plot(t2, np.cos(2*np.pi*t2), 'r--');
plt.show()
```



5.8 Resources

A really nice online resource for learning data visualization in Python is the Python Graph Gallery. This site has many examples of different kinds of plots using `pandas`, `seaborn` and `matplotlib`

```
reticulate::use_condaenv('ds', required=T)
```

6 Statistical analysis

6.1 Introduction

Statistical analysis usually encompasses 3 activities in a data science workflow. These are (a) descriptive analysis, (b) hypothesis testing and (c) statistical modeling. Descriptive analysis refers to a description of the data, which includes computing summary statistics and drawing plots. Hypothesis testing usually refers to statistically seeing if two (or more) groups are different from each other based on some metrics. Modeling refers to fitting a curve to the data to describe the relationship patterns of different variables in a data set.

In terms of Python packages that can address these three tasks:

Task	Packages
Descriptive statistics	pandas, numpy, matplotlib, seaborn
Hypothesis testing	scipy, statsmodels
Modeling	statsmodels, lifelines, scikit-learn

6.2 Descriptive statistics

Descriptive statistics that are often computed are the mean, median, standard deviation, inter-quartile range, pairwise correlations, and the like. Most of these functions are available in `numpy`, and hence are available in `pandas`. We have already seen how we can compute these statistics and have even computed grouped statistics. For example, we will compute these using the diamonds dataset

```
import numpy as np
import scipy as sc
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_context('paper')
sns.set_style('white', {'font.family': 'Future Medium'})
```

```
diamonds = pd.read_csv('data/diamonds.csv.gz')
```

```
diamonds.groupby('color')['price'].agg([np.mean, np.median, np.std])
```

```
              mean  median          std
```

```
color
D      3169.954096  1838.0  3356.590935
E      3076.752475  1739.0  3344.158685
F      3724.886397  2343.5  3784.992007
G      3999.135671  2242.0  4051.102846
H      4486.669196  3460.0  4215.944171
I      5091.874954  3730.0  4722.387604
J      5323.818020  4234.0  4438.187251
```

There were other examples we saw yesterday along these lines. Refer to both the `python_tools_ds` and `python_pandas` documents

6.3 Classical hypothesis testing

Python has the tools to do classic hypothesis testing. Several functions are available in the `scipy.stats` module. The commonly used tests that are available are as follows:

Function	Test
`ttest_1samp`	One-sample t-test
`ttest_ind`	Two-sample t-test
`ttest_rel`	Paired t-test
`wilcoxon`	Wilcoxon signed-rank test (nonparametric paired t-test)
`mannwhitneyu`	Wilcoxon rank-sum test (nonparametric 2-sample t-test)
`chi2_contingency`	Chi-square test for independence
`fisher_exact`	Fisher's exact test on a 2x2 contingency table
`f_oneway`	One-way ANOVA
`pearsonr`	Testing for correlation

There are also several tests in `statsmodels.stats`

Functions	Tests
`proportions_ztest`	Test for difference in proportions
`mcnemar`	McNemar's test
`sign_test`	Sign test
`multipletests`	p-value correction for multiple tests
`fdrcorrection`	p-value correction by FDR

Let us look at a breast cancer proteomics experiment to illustrate this. The experimental data contains protein expression for over 12 thousand proteins, along with clinical data. We can ask, for example, whether a particular protein expression differs by ER status.

```
brca = pd.read_csv('data/brca.csv')
brca.head()
```

```
   Unnamed: 0 Complete TCGA ID  Gender  ...  NP_004065 NP_068752 NP_219494
0           0     TCGA-A2-A0CM  FEMALE  ...        NaN       NaN       NaN
1           1     TCGA-BH-A18Q  FEMALE  ...  -1.778435       NaN -3.069752
2           2     TCGA-A7-A0CE  FEMALE  ...        NaN       NaN       NaN
3           3     TCGA-D8-A142  FEMALE  ...        NaN       NaN       NaN
4           4     TCGA-AO-A0J6  FEMALE  ...        NaN       NaN -3.753616

[5 rows x 12585 columns]
```

We will use both the t-test and the Wilcoxon rank-sum test, the nonparametric equivalent.

We will first do the classical t-test, that is available in the `scipy` package.

```
import scipy as sc
import statsmodels as sm
test_probe = 'NP_001193600'

tst = sc.stats.ttest_ind(brca[brca['ER Status']=='Positive'][test_probe], # Need [] sinc
                   brca[brca['ER Status']=='Negative'][test_probe],
                  nan_policy = 'omit')
np.round(tst.pvalue, 3)
```

```
0.277
```

We will now do the Wilcoxon test, also known as the Mann-Whitney U test.

```
tst = sc.stats.mannwhitneyu(brca[brca['ER Status']=='Positive'][test_probe], # Need [] s
                   brca[brca['ER Status']=='Negative'][test_probe],
                  alternative = 'two-sided')
np.round(tst.pvalue, 3)
```

```
0.996
```

We will come back to this when we look at permutation tests below.

6.4 Simulation and inference

Hypothesis testing is one of the areas where statistics is often used. There are functions for a lot of the standard statistical tests in `scipy` and `statsmodels`. However, I'm going to take a little detour to see if we can get some understanding of hypothesis tests using the powerful simulation capabilities of Python. We'll visit the in-built functions available in `scipy` and `statsmodels` as well.

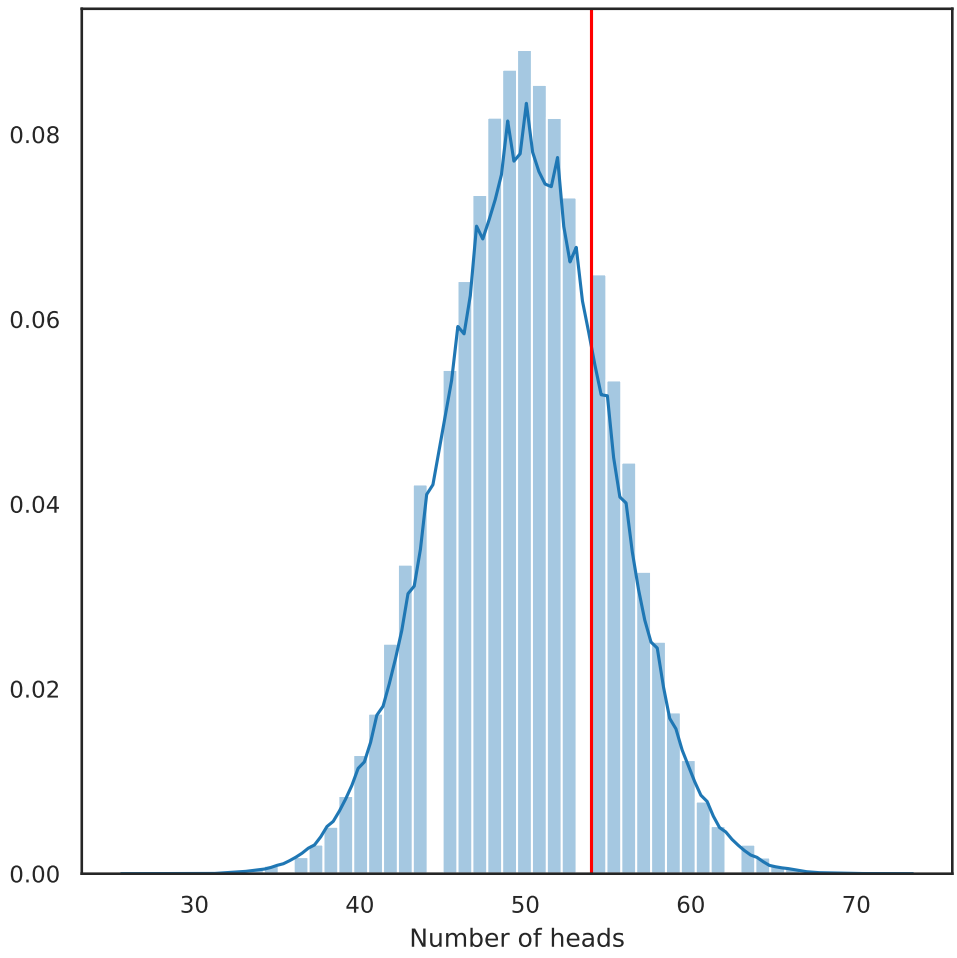
6.4.1 Simulation and hypothesis testing

Question: You have a coin and you flip it 100 times. You get 54 heads. How likely is it that you have a fair coin?

We can simulate this process, which is random, using Python. The process of heads and tails from coin tosses can be modeled as a **binomial** distribution. We can repeat this experiment many many times on our computer, making the assumption that we have a fair coin, and then seeing how likely what we observed is under that assumption.

> Simulation under reasonable assumptions is a great way to understand our data and the underlying data generating processes. In the modern era, it has most famously been used by Nate Silver of ESPN to simulate national elections in the US. There are many examples in engineering where simulations are done to understand a technology and figure out its tolerances and weaknesses, like in aircraft testing. It is also commonly used in epidemic modeling to help understand how an epidemic would spread under different conditions.

```
rng = np.random.RandomState(205) # Seed the random number generator

x = rng.binomial(100, 0.5, 100000) # Simulate 100,000 experiments of tossing a fair coin

sns.distplot(x, kde=True, rug=False)
plt.axvline(54, color = 'r'); # What we observed
plt.xlabel('Number of heads');
```




```
# We convert to pd.Series to take advantage of the `describe` function.
pd.Series(x).describe()
```

```
count    100000.000000
mean         49.995590
std           5.011938
min          27.000000
25%          47.000000
50%          50.000000
75%          53.000000
max          72.000000
dtype: float64
```

What we see from the histogram and the description of the data above is the patterns in data we would expect if we repeated this random experiment. We can already make some observations. First, we do see

that the average number of heads we expect to get is 50, which validates that our experiment is using a fair coin. Second, we can reasonably get as few as 27 heads and as many as 72 heads even with a fair coin. In fact, we could look at what values we would expect to see 95% of the time.

```
np.quantile(x, [0.025, 0.975])
```

```
array([40., 60.])
```

This says that 95% of the time we'll see values between 40 and 60. (This is **not** a confidence interval. This is the actual results of a simulation study. A confidence interval would be computed based on a **single** experiment, assuming a binomial distribution. We'll come to that later).

So how likely would we be to see the 54 heads in 100 tosses assuming a fair coin? This can be computed as the proportion of experiments

```
np.mean(x > 54) # convince yourself of this
```

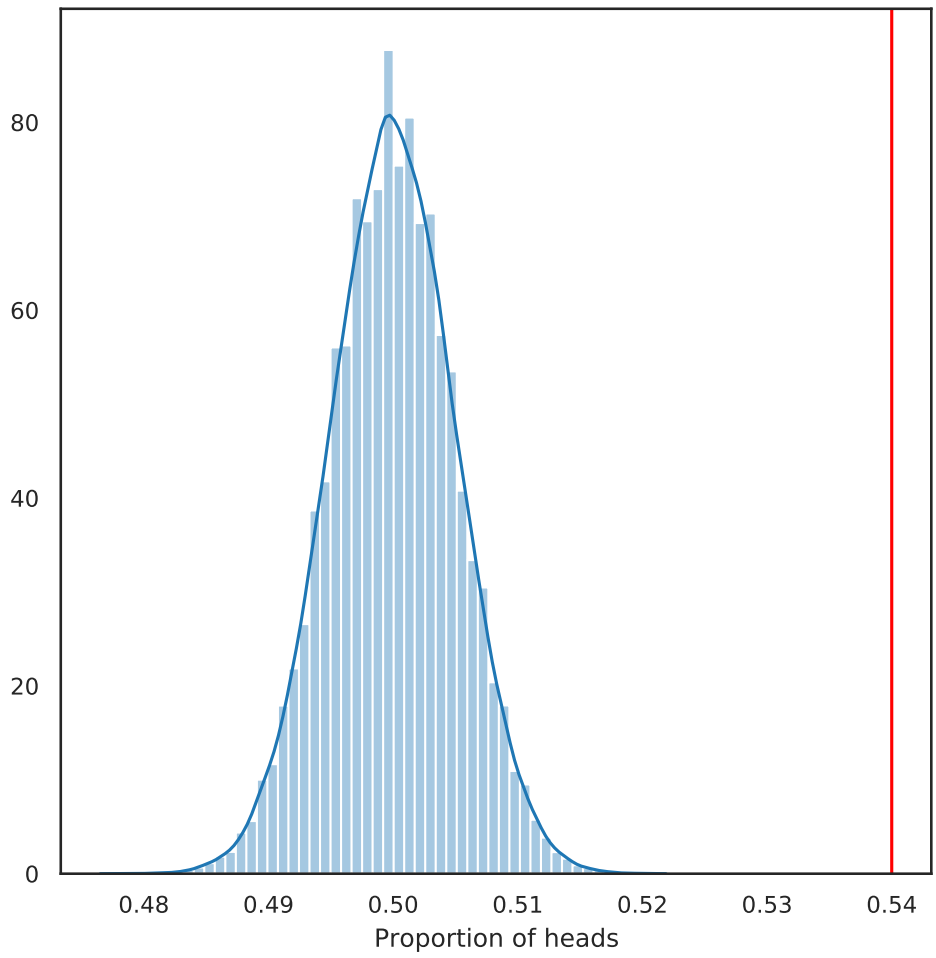
```
0.18456
```

This is what would be considered the *p-value* for the test that the coin is fair.

> The p-value of a statistical hypothesis test is the likelihood that we would see an outcome at least as extreme as we observed under the assumption that the null hypothesis (H0) that we chose is actually true.
>
> In our case, that null hypothesis is that the coin we're tossing is fair. The p-value **only** gives evidence against the null hypothesis, but does **not** give evidence for the null hypothesis. In other words, if the p-value is small (smaller than some threshold we deem reasonable), then we can claim evidence against the null hypothesis, but if the p-value is large, we cannot say the null hypothesis is true.

What happens if we increase the number of tosses, and we look at the proportion of heads. We observe 54% heads.

```
rng = np.random.RandomState(205)
x = rng.binomial(10000, 0.5, 100000)/10000
sns.distplot(x)
plt.axvline(0.54, color = 'r')
plt.xlabel('Proportion of heads');
```




```
pd.Series(x).describe()
```

```
count    100000.000000
mean          0.499991
std           0.004994
min           0.478100
25%           0.496600
50%           0.500000
75%           0.503400
max           0.520300
dtype: float64
```

Well, that changed the game significantly. If we up the number of coin tosses per experiment to 10,000, so 100-fold increase, then we do not see very much variation in the proportion of tosses that are heads.

> This is expected behavior because of a statistical theorem called the *Law of Large Numbers*, which essentially says that if you do larger and larger sized random experiments with the same experimental setup, your estimate of the true population parameter (in this case the true chance of getting a head, or 0.5 for a fair coin) will become more and more precise.

Now we see that for a fair coin, we should reasonably see between 47.8% and 52% of tosses should be heads. This is quite an improvement from the 27%-72% range we saw with 100 tosses.

We can compute our p-value in the same way as before.

```
np.mean(x > 0.54)
```

```
0.0
```

So we would never see 54% of our tosses be heads if we tossed a fair coin 10,000 times. Now, with a larger experiment, we would **reject** our null hypothesis H0 that we have a fair coin.

So same observation, but more data, changes our *inference* from not having sufficient evidence to say that the coin isn't fair to saying that it isn't fair quite definitively. This is directly due to the increased precision of our estimates and thus our ability to differentiate between much smaller differences in the truth.

Let's see a bit more about what's going on here. Suppose we assume that the coin's true likelihood of getting a head is really 0.55, so a very small bias towards heads.

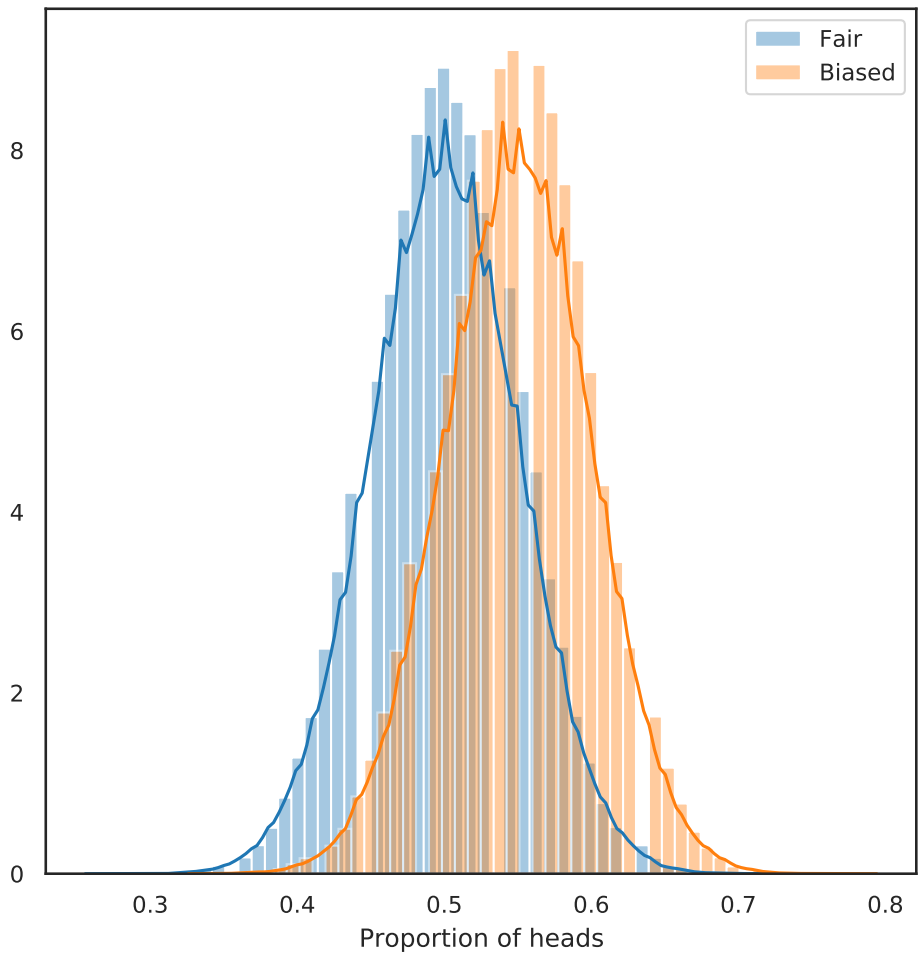
> Food for thought: Is the difference between 0.50 and 0.54 worth worrying about? It probably depends.

We're going to compare what we would reasonably see over many repeated experiments given the coin has a 0.50 (fair) and a 0.55 (slightly biased) chance of a head. First, we'll do experiments of 100 tosses of a coin.

```
rng = np.random.RandomState(205)
x11 = rng.binomial(100, 0.5, 100000)/100 # Getting proportion of heads
x12 = rng.binomial(100, 0.55, 100000)/100

sns.distplot(x11, label = 'Fair')
sns.distplot(x12, label = 'Biased')
```

```
<matplotlib.axes._subplots.AxesSubplot object at 0x13a81dcd0>
```

```
plt.xlabel('Proportion of heads')
plt.legend();
```



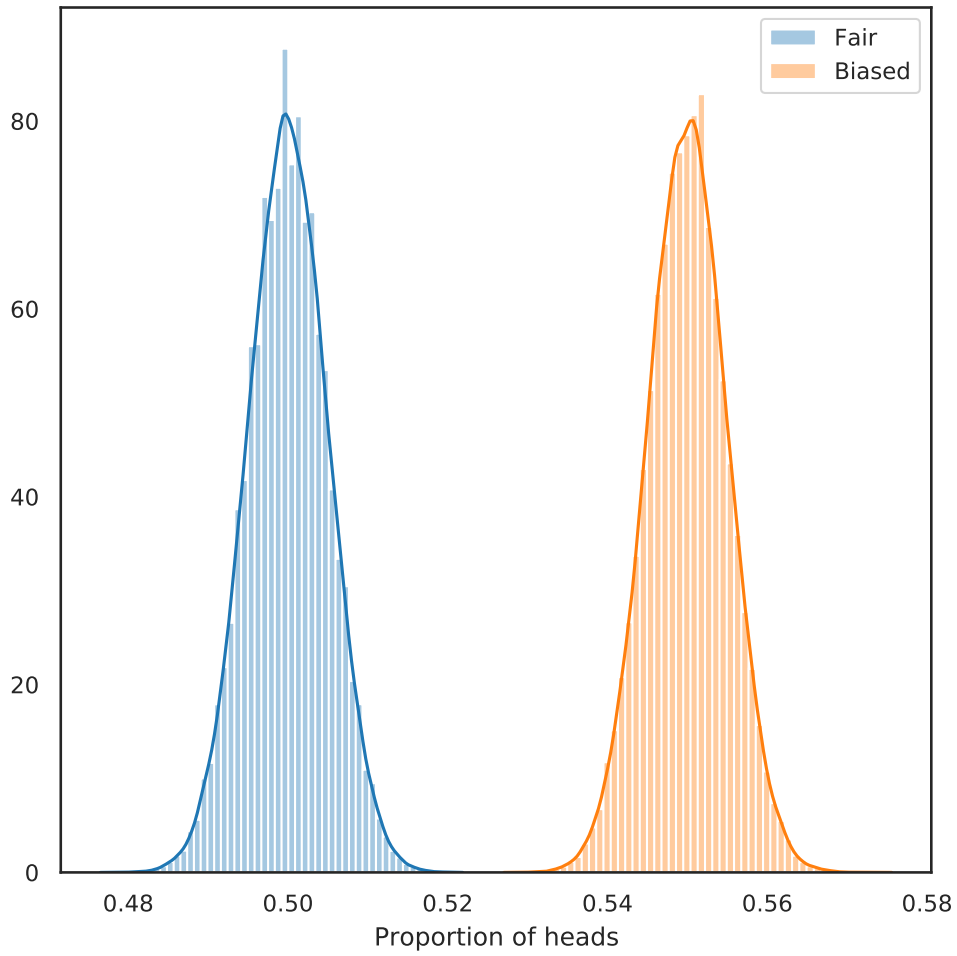
We see that there is a great deal of overlap in the potential outcomes over 100,000 repetitions of these experiments, so we have a lot of uncertainty about which model (fair or biased) is the truth.

Now, if we up our experiment to 10,000 tosses of each coin, and again repeat the experiment 100,000 times,

```
rng = np.random.RandomState(205)
x21 = rng.binomial(10000, 0.5, 100000)/10000
x22 = rng.binomial(10000, 0.55, 100000)/10000

sns.distplot(x21, label = 'Fair')
sns.distplot(x22, label = 'Biased')
```

```
<matplotlib.axes._subplots.AxesSubplot object at 0x13843c310>
```

```
plt.xlabel('Proportion of heads')
plt.legend();
```



We now find almost no overlap between the potential outcomes, so we can very easily distinguish the two models. This is part of what gathering more data (number of tosses) buys you.

We typically measure this ability to distinguish between two models using concepts of *statistical power*, which is the likelihood that we would find an observation at least as extreme as what we observed, under the **alternative** model (in this case, the biased coin model). We can calculate the statistical power quite easily for the two sets of simulated experiments. Remember, we observed 54% heads in our one instance of each experiment that we actually observed. By doing simulations, we're "playing God" and seeing what could have happened, but in practice we only do the experiment once (how many clinical trials of an expensive drug would you really want to do?).

```
pval1 = np.mean(x11 > 0.54)
pval2 = np.mean(x21 > 0.54)

power1 = np.mean(x12 > 0.54)
power2 = np.mean(x22 > 0.54)

print('The p-value when n=100 is ', np.round(pval1, 2))
```

```
The p-value when n=100 is  0.18
```

```
print('The p-value when n=10,000 is ', np.round(pval2, 2))
```

```
The p-value when n=10,000 is  0.0
```

```
print('Statistical power when n=100 is ', np.round(power1, 2))
```

```
Statistical power when n=100 is  0.54
```

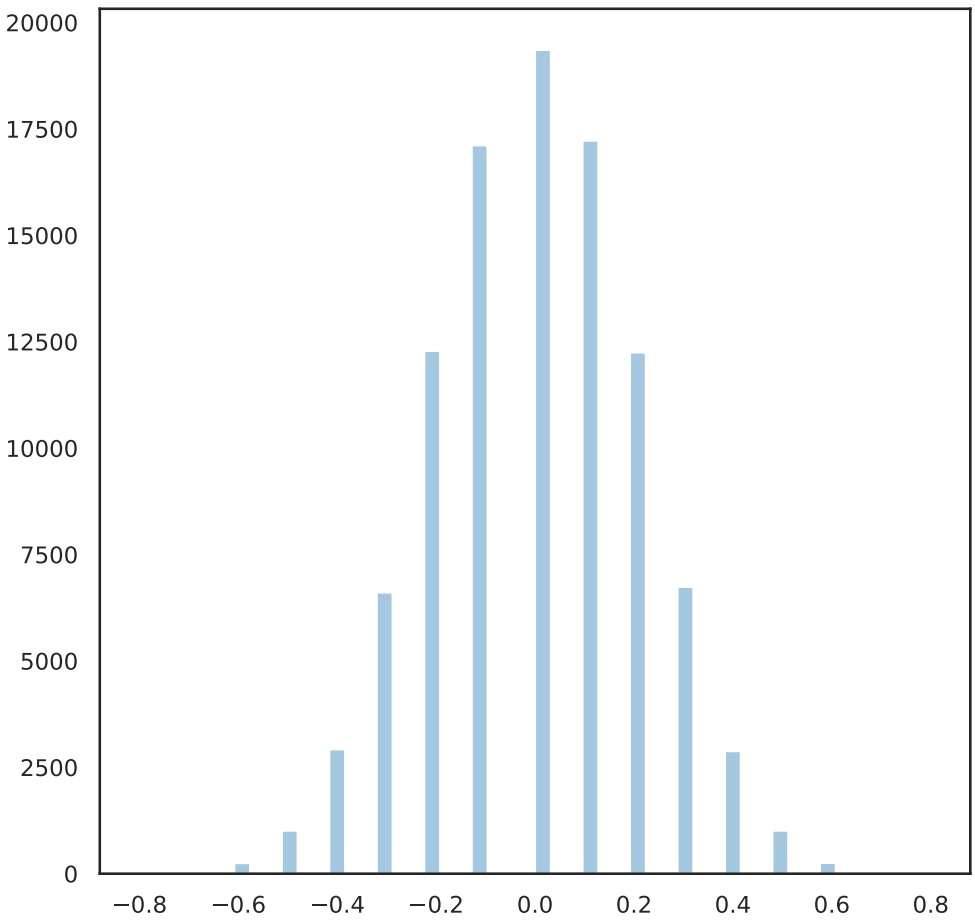
```
print('Statistical power when n=10,000 is ', np.round(power2, 2))
```

```
Statistical power when n=10,000 is  0.98
```

So as n goes up, the p-value for the same experimental outcome goes down and the statistical power goes up. This is a general rule with increasing sample size.

This idea can be used to design a two-armed experiment. Suppose we are looking at the difference in proportion of mice who gained weight between a wild-type mouse and a knockout variant. Since mice are expensive, let's limit the number of mice we'll use in each arm to 10. We expect 30% of the wild-type mice to gain weight, and expect a higher proportion of the knockouts will gain weight. This is again the setup for a binomial experiment, with the number of "coin tosses" being 10 for each of the arms. We're going to do two sets of experiments, one for the WT and one for the KO, and see the difference in proportions of weight gain ('heads') between them, and repeat it 100,000 times.

```
rng = np.random.RandomState(304)
N = 10
weight_gain_wt0 = rng.binomial(N, 0.3, 100000)/N # Get proportion
weight_gain_ko0 = rng.binomial(N, 0.3, 100000)/N # Assume first (null hypothesis) that t

diff_weight_gain0 = weight_gain_ko0 - weight_gain_wt0
sns.distplot(diff_weight_gain0, kde=False); # Since we only have 10 mice each, this hist
                                            # No matter!
```



We usually design the actual test by choosing a cutoff in the difference in proportions and stating that we will reject the null hypothesis if our observed difference exceeds this cutoff. We choose the cutoff so that the p-value of the cutoff is some pre-determined error rate, typically 0.05 or 5% (This is not golden or set in stone. We'll discuss this later). Let's find that cutoff from this simulation. This will correspond to the 95th percentile of this simulated distribution.

```
np.round(np.quantile(diff_weight_gain0, 0.95), 2)
```

```
0.3
```

This means that at least 5% of the values will be 0.3 or bigger. In fact, this proportion is

```
np.mean(diff_weight_gain0 > 0.3)
```

```
0.06673
```

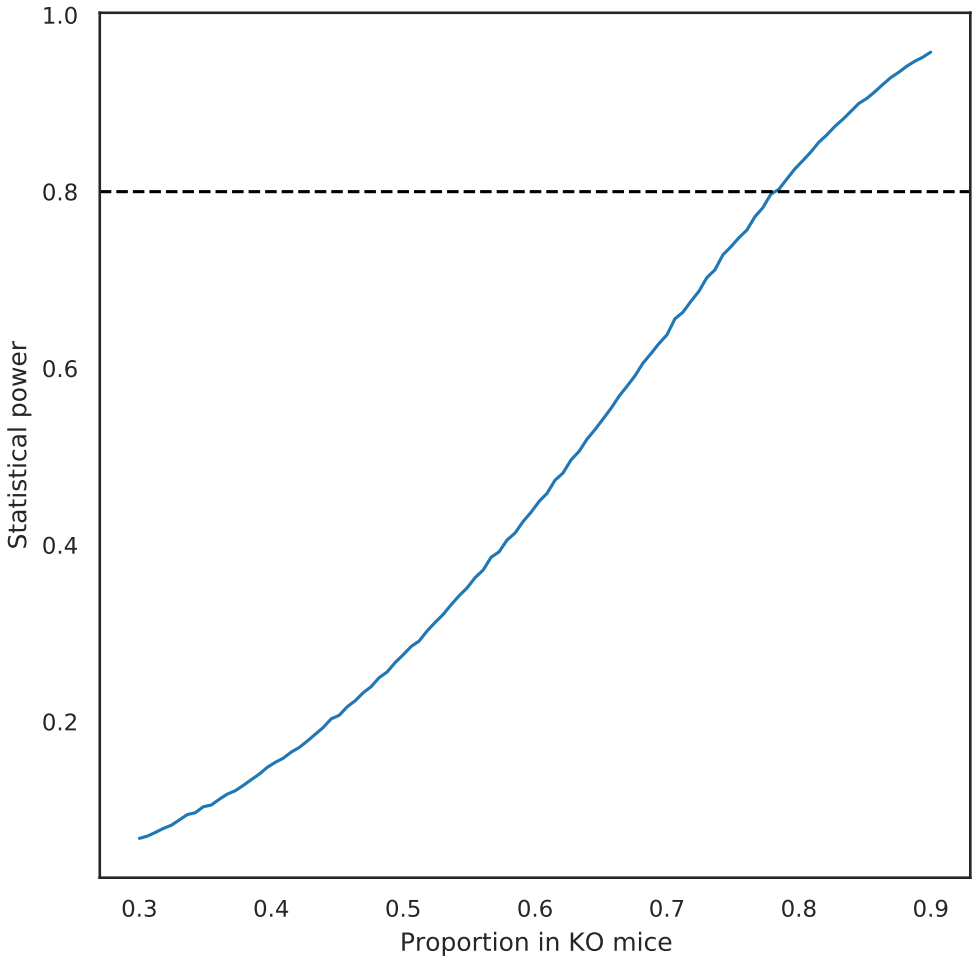
So we'll take 0.3 as the cutoff for our test (It's fine if the Type 1 error is more than 0.05. If we take the next largest value in the simulation, we dip below 0.05). We're basically done specifying the testing rule.

What we (and reviewers) like to know at this point is, what is the difference level for which you might get 80% power. The thinking is that if the true difference was, say, $p > 0$ rather than 0 (under the null hypothesis), we would reject the null hypothesis, i.e., get our observed difference to be more than 0.3, at least 80% of the time. We want to find out how big that value of p is. In other words, what is the level of difference in proportions at which we can be reasonably certain that our test will REJECT H0, given our sample size, when the true difference in proportions is p. Another way of saying this is how big does the difference in true proportions have to be before we would be fairly confident statistically of distinguishing that we have a difference between the two groups given our chosen sample size, i.e., fairly small overlaps in the two competing distributions.

We can also do this using simulation, by keeping the WT group at 0.3, increasing the KO group gradually, simulating the distribution of the difference in proportion and seeing at what point we get to a statistical power of about 80%. Recall, we've already determined that our test will reject H0 when the observed difference is greater than 0.3

```
p1 = np.linspace(0.3, 0.9, 100)
power = np.zeros(len(p1))
for i, p in enumerate(p1):
    weight_gain_wt1 = rng.binomial(N, 0.3, 100000)/N
    weight_gain_ko1 = rng.binomial(N, p, 100000)/N
    diff_weight_gain1 = weight_gain_ko1 - weight_gain_wt1
    power[i] = np.mean(diff_weight_gain1 > 0.3)
```

```
sns.lineplot(p1, power)
plt.axhline(0.8, color = 'black', linestyle = '--');
plt.ylabel('Statistical power')
plt.xlabel('Proportion in KO mice');
```




```
np.round(p1[np.argmin(np.abs(power - 0.8))] - 0.3, 2) # Find the location in the p1 arra
```

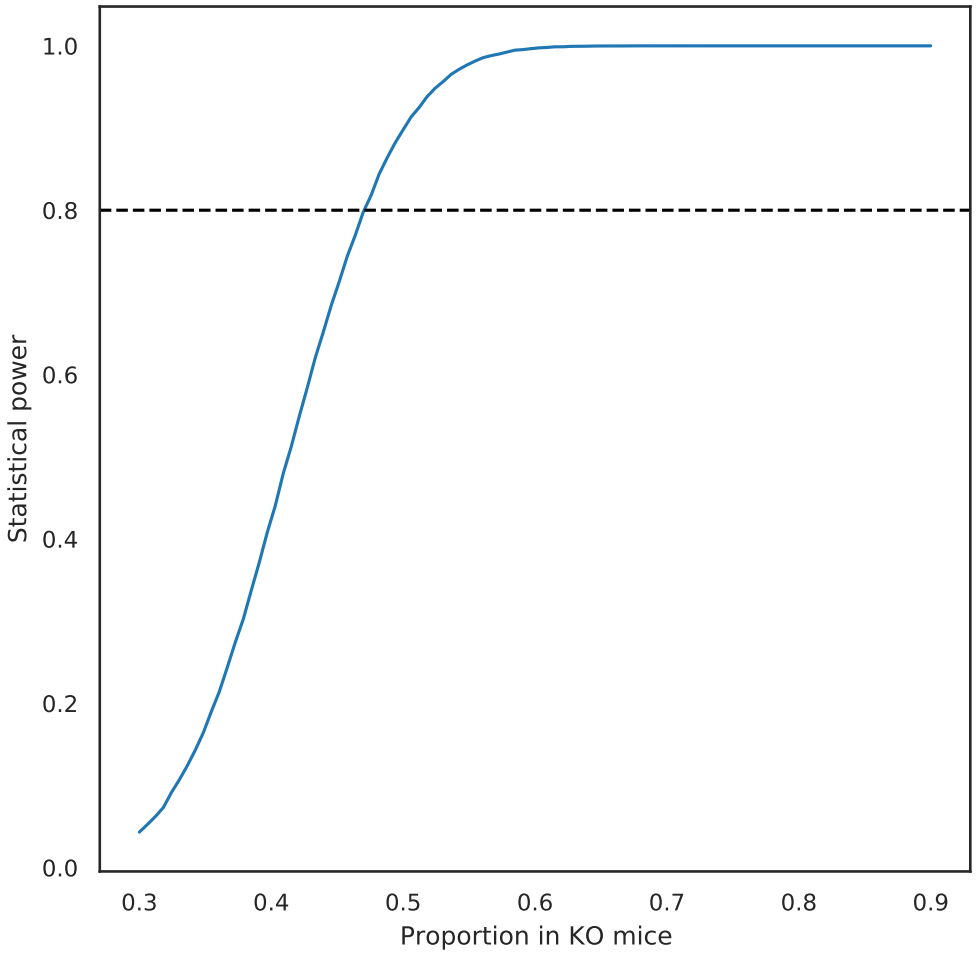
```
0.48
```

So to get to 80% power, we would need the true difference in proportion to be 0.48, or that at least 78% of KO mice should gain weight on average. This is quite a big difference, and its probably not very interesting scientifically to look for such a big difference, since it's quite unlikely.

If we could afford 100 mice per arm, what would this look like?

```
rng = np.random.RandomState(304)
N = 100
weight_gain_wt0 = rng.binomial(N, 0.3, 100000)/N # Get proportion
weight_gain_ko0 = rng.binomial(N, 0.3, 100000)/N # Assume first (null hypothesis) that t
```

```
diff_weight_gain0 = weight_gain_ko0 - weight_gain_wt0
cutoff = np.quantile(diff_weight_gain0, 0.95)

p1 = np.linspace(0.3, 0.9, 100)
power = np.zeros(len(p1))
for i, p in enumerate(p1):
    weight_gain_wt1 = rng.binomial(N, 0.3, 100000)/N
    weight_gain_ko1 = rng.binomial(N, p, 100000)/N
    diff_weight_gain1 = weight_gain_ko1 - weight_gain_wt1
    power[i] = np.mean(diff_weight_gain1 > cutoff)

sns.lineplot(p1, power)
plt.axhline(0.8, color = 'black', linestyle = '--');
plt.ylabel('Statistical power')
plt.xlabel('Proportion in KO mice');
```



```
np.round(p1[np.argmin(np.abs(power - 0.8))] - 0.3, 2)
```

```
0.17
```

The minimum detectable difference for 80% power is now down to 0.17, so we'd need the KO mice in truth to show weight gain 47% of the time, compared to 30% in WT mice. This is more reasonable scientifically as a query.

6.4.2 A permutation test

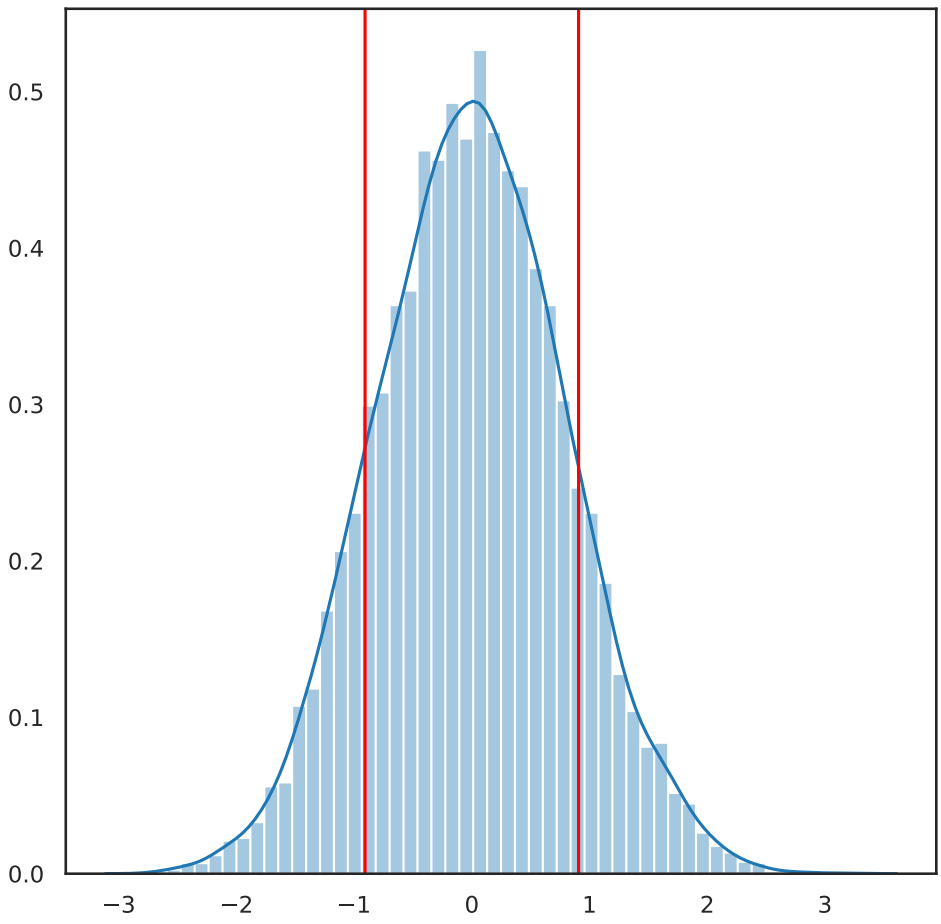
A permutation test is a 2-group test that asks whether two groups are different with respect to some metric. We'll use the same proteomic data set as before.

The idea about a permutation test is that, if there is truly no difference then it shouldn't make a difference if we shuffled the labels of ER status over the study individuals. That's literally what we will do. We will do this several times, and look at the average difference in expression each time. This will form the null distribution under our assumption of no differences by ER status. We'll then see where our observed data falls, and then be able to compute a p-value.

The difference between the simulations we just did and a permutation test is that the permutation test is based only on the observed data. No particular models are assumed and no new data is simulated. All we're doing is shuffling the labels among the subjects, but keeping their actual data intact.

```
nsim = 10000

rng = np.random.RandomState(294)
x = np.where(brca['ER Status']=='Positive', -1, 1)
y = brca[test_probe].to_numpy()

obs_diff = np.nanmean(y[x==1]) - np.nanmean(y[x==-1])

diffs = np.zeros(nsim)
for i in range(nsim):
    x1 = rng.permutation(x)
    diffs[i] = np.nanmean(y[x1==1]) - np.nanmean(y[x1 == -1])
```

```
sns.distplot(diffs)
plt.axvline(x = obs_diff, color ='r');
plt.axvline(x = -obs_diff, color = 'r');
```




```
pval = np.mean(np.abs(diffs) > np.abs(obs_diff))
f"P-value from permutation test is {pval}"
```

```
'P-value from permutation test is 0.2606'
```

This is pretty close to what we got from the t-test.

Note that what we've done here is the two-sided test to see how extreme our observation would be in either direction. That is why we've taken the absolute values above, and drawn both the observed value and it's negative on the graph.

6.4.3 Testing many proteins

We could do the permutation test all the proteins using the array operations in numpy

```
expr_names = [u for u in list(brca.columns) if u.find('NP') > -1]
             # Find all column names with NP

exprs = brca[expr_names] # Extract the protein data
```

```
x = np.where(brca['ER Status']=='Positive', -1, 1)
obs_diffs = exprs[x==1].mean(axis=0)-exprs[x==-1].mean(axis=0)
```

```
nsim = 1000
diffs = np.zeros((nsim, exprs.shape[1]))
for i in range(nsim):
    x1 = rng.permutation(x)
    diffs[i,:] =exprs[x1==1].mean(axis=0) - exprs[x1==-1].mean(axis=0)
```

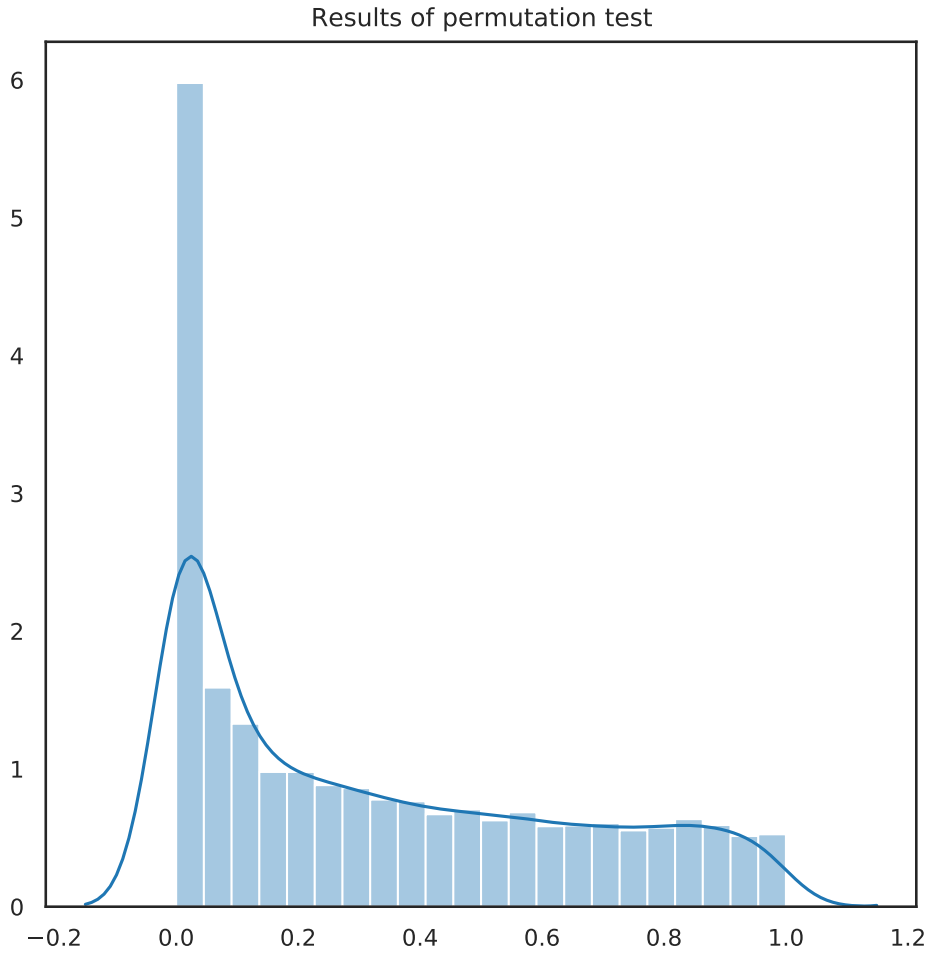
```
pvals = np.zeros(exprs.shape[1])
len(pvals)
```

```
12395
```

```
for i in range(len(pvals)):
    pvals[i] = np.mean(np.abs(diffs[:,i]) > np.abs(obs_diffs.iloc[i]))
```

```
<string>:2: RuntimeWarning: invalid value encountered in greater
```

```
sns.distplot(pvals);
plt.title('Results of permutation test')
```



This plot shows that there is probably some proteins which are differentially expressed between ER+ and ER- patients. (If no proteins had any difference, this histogram would be flat, since the p-values would be uniformly distributed). The ideas around Gene Set Enrichment Analysis (GSEA) can also be applied here.

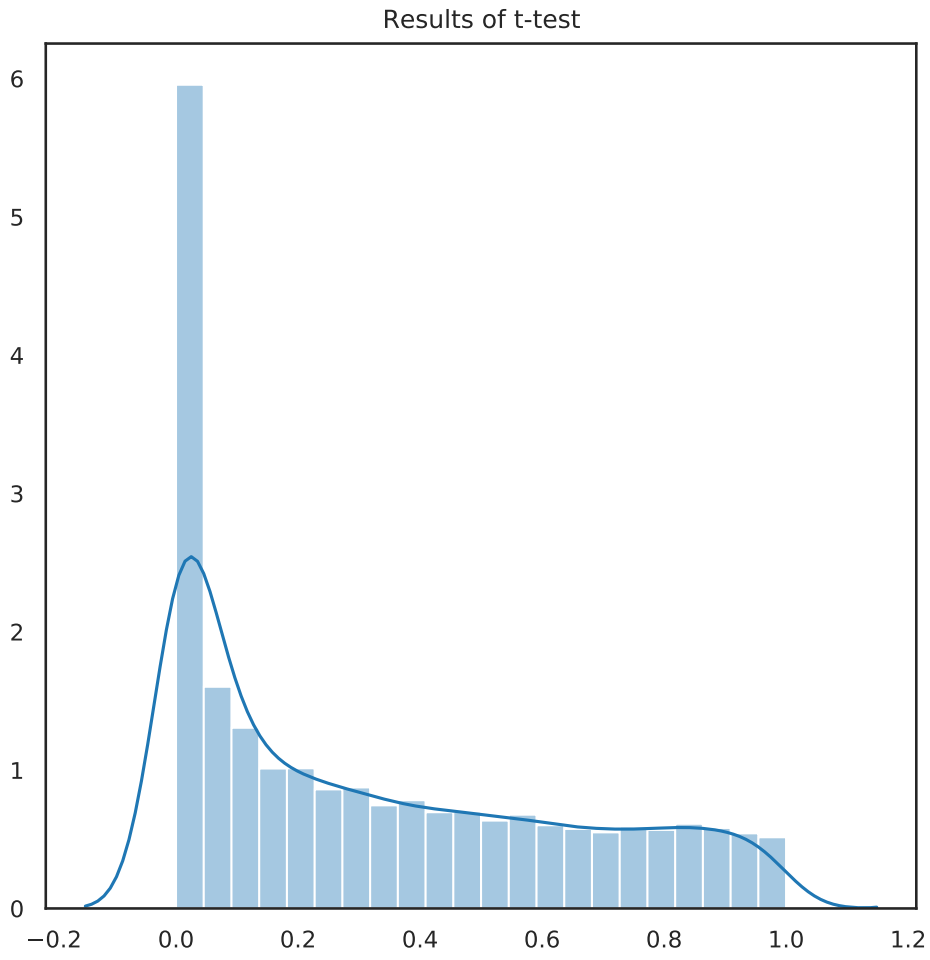
```
exprs_shortlist = [u for i, u in enumerate(list(exprs.columns))
                   if pvals[i] < 0.0001 ]

len(exprs_shortlist)
```

```
896
```

This means that, if we considered a p-value cutoff for screening at 0.0001, we would select 896 of the 12395 proteins for further study. Note that if none of the proteins had any effect, we'd expect 0.0001 x 12395 or 13 proteins to have a p-value smaller than 0.0001.

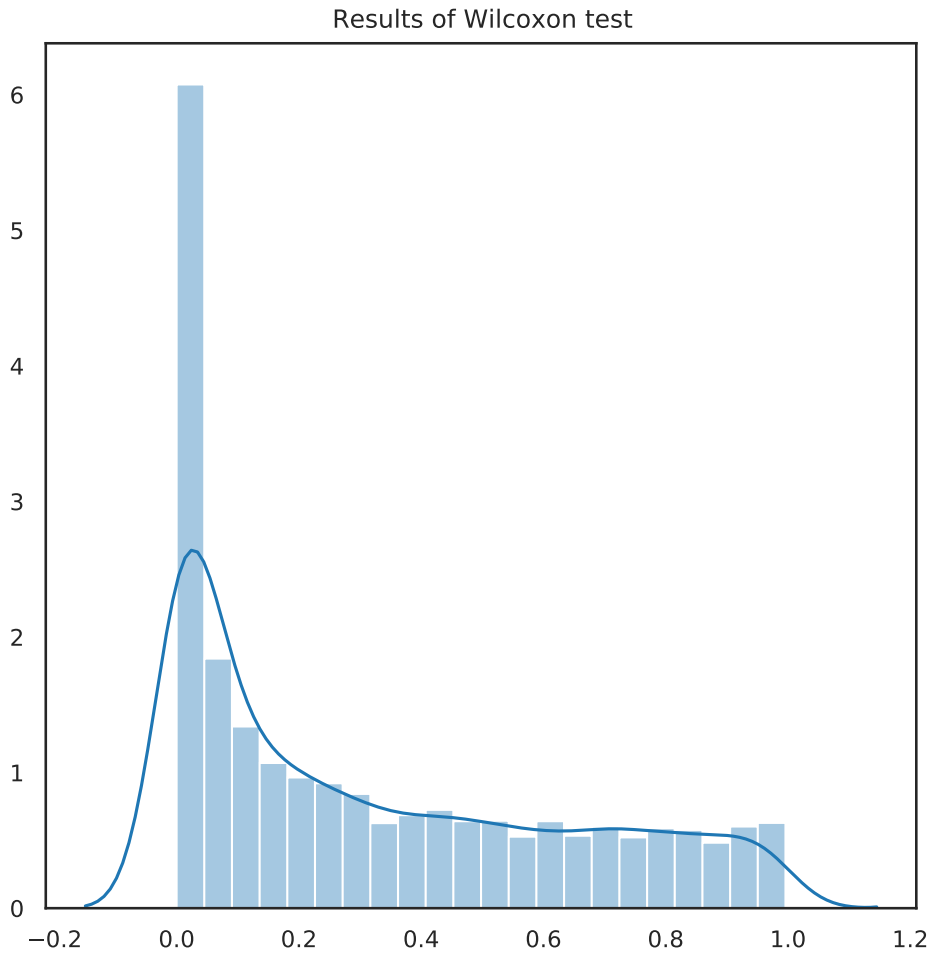
We could also do the same thing using both the t-test and the Mann-Whitney test.

```
groups = np.where(brca['ER Status']=='Positive', 1, 0)
pvals_t = np.zeros(exprs.shape[1])
for i in range(exprs.shape[1]):
    stat, pvals_t[i] = sc.stats.ttest_ind(exprs.iloc[groups==1, i],
                                          exprs.iloc[groups==0, i],
                                          nan_policy = 'omit')
sns.distplot(pvals_t);
plt.title('Results of t-test');
```



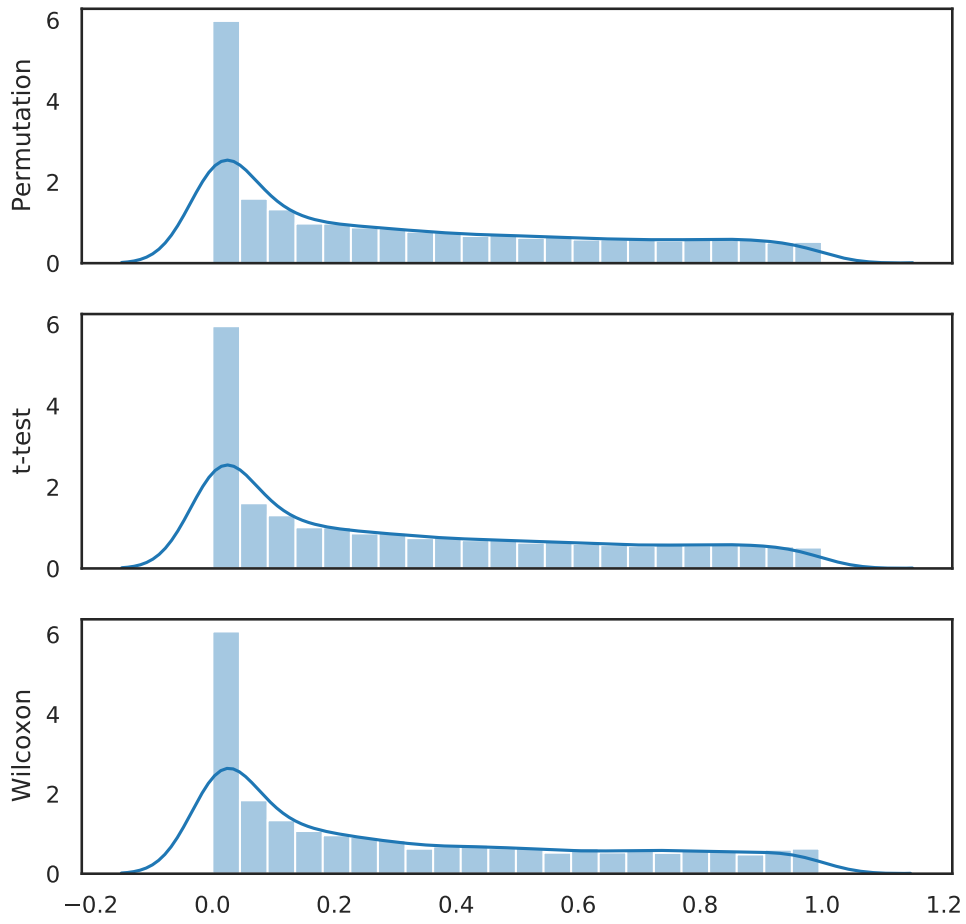

```
pvals_w = np.zeros(exprs.shape[1])
for i in range(exprs.shape[1]):
    stats, pvals_w[i] = sc.stats.mannwhitneyu(exprs.iloc[groups==1,i],
                                              exprs.iloc[groups==0, i],
```

```
                                                alternative='two-sided')
sns.distplot(pvals_w);
plt.title('Results of Wilcoxon test');
```



We can directly compare the graphs, which appear quite similar.

```
fig, ax = plt.subplots(3,1, sharex = True)

sns.distplot(pvals, ax = ax[0]); ax[0].set_ylabel('Permutation');
sns.distplot(pvals_t, ax = ax[1]); ax[1].set_ylabel('t-test');
sns.distplot(pvals_w, ax = ax[2]); ax[2].set_ylabel('Wilcoxon');
```



We can also compare how many proteins will be chosen if we employ a p-value cutoff of 0.0001

```
pvalues = pd.DataFrame({'permutation' : pvals, 'ttest' : pvals_t,
                        'wilcoxon' : pvals_w})
pvalues.apply(lambda x: np.sum(x < 0.0001))
```

```
permutation    896
ttest          499
wilcoxon       396
dtype: int64
```

The **lambda function** employed above is an anonymous (un-named) function that can be used on-the-fly. In the above statement, this function takes one (vector) argument *x* and computes the number of *x* values less than 0.0001. This function is then applied to each column of the `pvalues` dataset using the `apply` function.

6.4.4 Getting a confidence interval using the bootstrap

We can use simulations to obtain a model-free confidence interval for particular parameters of interest based on our observed data. The technique we will demonstrate is called the bootstrap. The idea is that if we sample with replacement from our observed data to get another data set of the same size as the observed data, and compute our statistic of interest, and then repeat this process many times, then the distribution of our statistic that we will obtain this way will be very similar to the true sampling distribution of the statistic if we could "play God". This has strong theoretical foundations from work done by several researchers in the 80s and 90s.

1. Choose the number of simulations `nsim`
2. for each iteration (1,...,nsim)
    - Simulate a dataset with replacement from the original data.
    - compute and store the statistic
3. Compute the 2.5th and 97.5th percential of the distribution of the statistic. This is your confidence interval.

Let's see this in action. Suppose we tossed a coin 100 times. We're going to find a confidence interval for the proportion of heads from this coin.

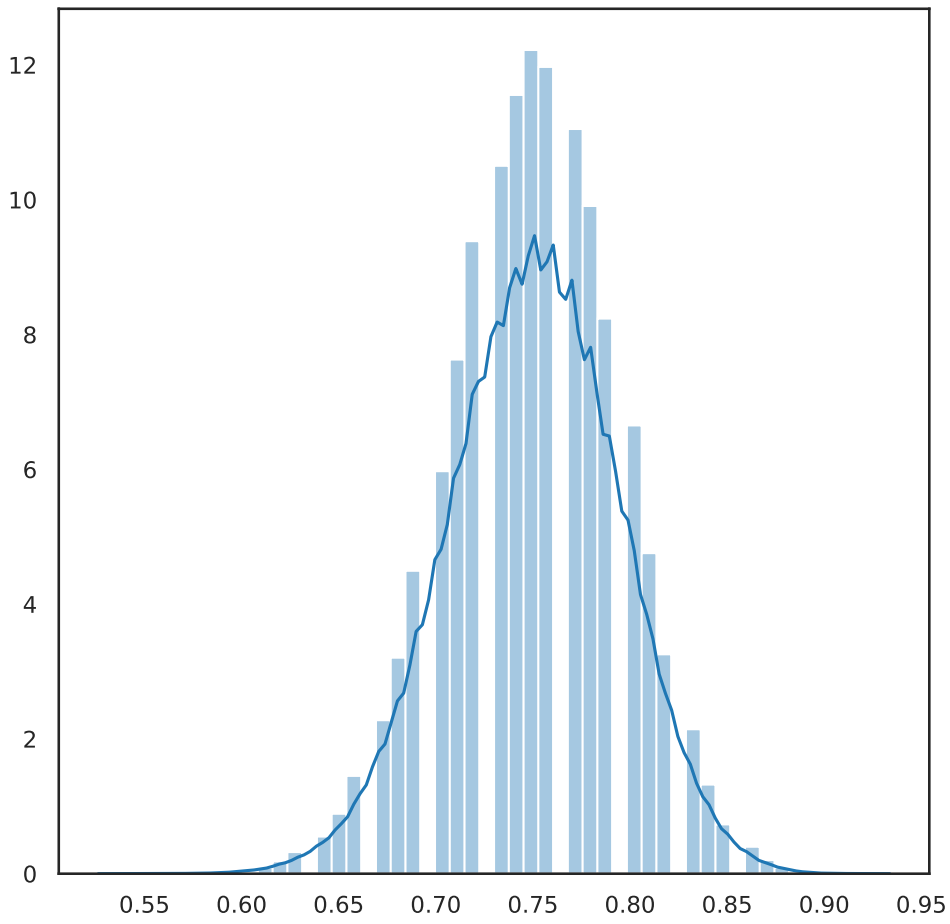
```
rng = np.random.RandomState(304)
x = rng.binomial(1, 0.7, 100)
x
```

```
array([1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0,
       1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0,
       1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1,
       1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1,
       1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1])
```

This gives the sequence of heads (1) and tails (0), assuming the true probability of heads is 0.7.

We now create 100000 bootstrap samples from here.

```
nsim = 100000

boots = np.random.choice(x, (len(x), nsim), replace = True) # sample from the data
boot_estimates = boots.mean(axis = 0) # compute mean of each sample, i.e proportion of h

sns.distplot(boot_estimates);
```




```
np.quantile(boot_estimates, (0.025, 0.975)) # Find 2.5 and 97.5-th percentiles
```

```
array([0.66, 0.83])
```

So our 95% bootstrap confidence interval is (0.66, 0.83). Our true value of 0.7 certainly falls in it.

6.5 Regression analysis

6.5.1 Ordinary least squares (linear) regression

The regression modeling frameworks in Python are mainly in `statsmodels`, though some of it can be found in `scikit-learn` which we will see tomorrow. We will use the diamonds dataset for demonstration purposes. We will attempt to model the diamond price against several of the other diamond characteristics.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf # Use the formula interface to statsmodels
```

```
diamonds = sm.datasets.get_rdataset('diamonds','ggplot2').data
mod1 = smf.ols('price ~ np.log(carat) + clarity + depth + cut * color', data = diamonds)
```

```
mod1.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.786
Method:                 Least Squares   F-statistic:                     4598.
Date:                Thu, 04 Jun 2020   Prob (F-statistic):               0.00
Time:                        18:55:08   Log-Likelihood:            -4.8222e+05
No. Observations:               53940   AIC:                         9.645e+05
Df Residuals:                   53896   BIC:                         9.649e+05
Df Model:                          43
Covariance Type:            nonrobust
================================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
--------------------
Intercept               2745.0643    415.804      6.602      0.000    1930.085    3560.043
clarity[T.IF]           4916.7221     83.694     58.746      0.000    4752.681    5080.763
clarity[T.SI1]          2686.1493     71.397     37.623      0.000    2546.210    2826.088
clarity[T.SI2]          2060.8180     71.809     28.699      0.000    1920.072    2201.564
clarity[T.VS1]          3710.1759     72.891     50.900      0.000    3567.309    3853.043
clarity[T.VS2]          3438.3999     71.792     47.894      0.000    3297.687    3579.112
clarity[T.VVS1]         4540.1420     77.314     58.724      0.000    4388.606    4691.678
clarity[T.VVS2]         4343.0545     75.136     57.803      0.000    4195.788    4490.321
cut[T.Good]              708.5981    161.869      4.378      0.000     391.334    1025.862
cut[T.Ideal]            1198.2067    149.690      8.005      0.000     904.812    1491.601
cut[T.Premium]          1147.1417    152.896      7.503      0.000     847.464    1446.820
cut[T.Very Good]        1011.3463    152.977      6.611      0.000     711.510    1311.183
color[T.E]                -59.4094    190.227     -0.312      0.755    -
432.256     313.437
color[T.F]                -86.0097    178.663     -0.481      0.630    -
436.191     264.172
color[T.G]               -370.6455    178.642     -2.075      0.038    -
720.784     -20.507
```

```
color[T.H]                    -591.0922    179.786     -3.288      0.001    -
943.474    -238.710
color[T.I]                   -1030.7417    201.485     -5.116      0.000    -
1425.655    -635.829
color[T.J]                   -1210.6501    223.111     -5.426      0.000    -
1647.949    -773.351
cut[T.Good]:color[T.E]         -30.3553    212.126     -0.143      0.886    -
446.123     385.413
cut[T.Ideal]:color[T.E]       -211.3711    195.630     -1.080      0.280    -
594.807     172.065
cut[T.Premium]:color[T.E]      -91.3261    199.440     -0.458      0.647    -
482.230     299.578
cut[T.Very Good]:color[T.E]    -45.2968    199.656     -0.227      0.821    -
436.625     346.031
cut[T.Good]:color[T.F]        -365.4060    202.035     -1.809      0.071    -
761.397      30.585
cut[T.Ideal]:color[T.F]       -198.0428    184.498     -1.073      0.283    -
559.661     163.575
cut[T.Premium]:color[T.F]     -322.8527    188.465     -1.713      0.087    -
692.246      46.540
cut[T.Very Good]:color[T.F]   -186.0519    189.090     -0.984      0.325    -
556.670     184.566
cut[T.Good]:color[T.G]         -93.0430    202.404     -0.460      0.646    -
489.757     303.671
cut[T.Ideal]:color[T.G]        -65.8579    183.980     -0.358      0.720    -
426.461     294.745
cut[T.Premium]:color[T.G]       35.4302    187.596      0.189      0.850    -
332.260     403.121
cut[T.Very Good]:color[T.G]    -81.2595    188.786     -0.430      0.667    -
451.282     288.764
cut[T.Good]:color[T.H]         137.0235    205.696      0.666      0.505    -
266.142     540.189
cut[T.Ideal]:color[T.H]        -83.4763    186.060     -0.449      0.654    -
448.155     281.202
cut[T.Premium]:color[T.H]      -44.4372    189.378     -0.235      0.814    -
415.620     326.745
cut[T.Very Good]:color[T.H]    -43.2485    190.851     -0.227      0.821    -
417.318     330.821
cut[T.Good]:color[T.I]         331.4048    228.614      1.450      0.147    -
116.681     779.490
cut[T.Ideal]:color[T.I]        106.2368    208.391      0.510      0.610    -
302.210     514.684
cut[T.Premium]:color[T.I]      357.1453    212.341      1.682      0.093     -
59.045     773.335
cut[T.Very Good]:color[T.I]    149.1555    213.697      0.698      0.485    -
269.693     568.004
cut[T.Good]:color[T.J]        -406.8484    256.938     -1.583      0.113    -
910.448      96.752
```

```
cut[T.Ideal]:color[T.J]      -330.0602    234.063     -1.410       0.159    -
788.826     128.706
cut[T.Premium]:color[T.J]    -156.8065    236.860     -0.662       0.508    -
621.055     307.442
cut[T.Very Good]:color[T.J]  -381.5722    238.799     -1.598       0.110    -
849.620      86.475
np.log(carat)           6630.7799    15.605   424.923    0.000   6600.195   6661.365
depth                     -0.7353     5.961    -0.123     0.902     -
12.418      10.948
==============================================================================
Omnibus:                    13993.592   Durbin-Watson:                   0.134
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            34739.732
Skew:                           1.432   Prob(JB):                         0.00
Kurtosis:                       5.693   Cond. No.                     7.08e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.08e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

This is the basic syntax for modeling in statsmodels using the *formula* interface. This formula interface mimics the way regression formula are written in R. We will use this formula interface here since it allows for a more concise expression of the regression formula, and handles several things, as we will see.

> `statsmodels` provides a traditional input syntax as well, where you specify the dependent or *endogenous* variable y as a vector array, and the independent or *exogenous* variables X as a numerical matrix. The typical syntax would be `mod2 = sm.OLS(y, X).fit()`. The formula interface, which uses the Python package **patsy**, takes care of the conversions, as well as modifying the design matrix to accommodate interactions and transformations.

Let's go through and parse it.

One thing you notice is that we've written a formula inside the model

```
mod1 = smf.glm('price ~ np.log(carat) + clarity + depth + cut * color',
    data = diamonds).fit()
```

This formula will read as "price depends on log(carat), clarity, depth, cut and color, and the interaction of cut and color". Underneath a lot is going on.

1. color, clarity, and cut are all categorical variables. They actually need to be expanded into dummy variables, so we will have one column for each category level, which is 1 when the diamond is of that category and 0 otherwise. We typically use the **treatment** contrast formulation, which deems one category (usually the first) to be the reference category, and so creates one less dummy variable than the number of category levels, corresponding to the reference level.
2. An intercept term is added

3. The variable `carat` is transformed using `np.log`, i.e. the natural logarithm available in the numpy package. Generally, any valid Python function can be used here, even ones you create.
4. Interactions are computed. The syntax `cut * color` is a shortcut for `cut + color + cut:color`, where the `:` denotes interaction.
5. The dummy variables are concatenated to the continuous variables
6. The model is run

To see the full design matrix we can drop down and use **patsy** functions:

```
import patsy
f = mod1.model.formula
y,X = patsy.dmatrices(f, data = diamonds, return_type = 'dataframe')
```

X is the full design matrix with all the transformations and dummy variables and interactions computed, as specified by the formula.

Suppose we wanted the Ideal cut of diamond to be the reference level for the `cut` variable. We could specify this within the formula quite simply as:

```
mod2 = smf.ols('price ~ np.log(carat) + clarity + depth + C(cut, Treatment("Ideal")) * c
```

This syntax says that we consider `cut` to be a categorical variable, from which we will create dummy variables using *treatment* contrasts, using Ideal as the reference level.

6.5.2 Logistic regression

Logistic regression is the usual regression method used when you have binary outcomes, e.g., Yes/No, Negative/Positive, etc.

Logistic regression does exist as an individual method in **scikit-learn**, whic we will see in the Machine Learning module. However, it resides in its more traditional form within the *generalized linear model* framework in **statsmodels**

We will use a dataset based on deaths from the Titanic disaster in 1912.

```
titanic = sm.datasets.get_rdataset('Titanic','Stat2Data').data
titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1313 entries, 0 to 1312
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Name      1313 non-null   object
 1   PClass    1313 non-null   object
 2   Age       756 non-null    float64
 3   Sex       1313 non-null   object
```

```
 4   Survived  1313 non-null   int64
 5   SexCode   1313 non-null   int64
dtypes: float64(1), int64(2), object(3)
memory usage: 61.7+ KB
```

We will model `Survived` on the age, sex and passenger class of passengers.

```
mod_logistic = smf.glm('Survived ~ Age + Sex + PClass', data=titanic,
  family = sm.families.Binomial()).fit()
mod_logistic.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:               Survived   No. Observations:                  756
Model:                            GLM   Df Residuals:                      751
Model Family:                Binomial   Df Model:                            4
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -347.57
Date:                Thu, 04 Jun 2020   Deviance:                       695.14
Time:                        18:55:10   Pearson chi2:                     813.
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
-----
Intercept        1.8664      0.217      8.587      0.000       1.440       2.292
Sex[T.male]     -2.6314      0.202    -13.058      0.000      -3.026      -2.236
PClass[T.1st]    1.8933      0.208      9.119      0.000       1.486       2.300
PClass[T.2nd]    0.6013      0.148      4.052      0.000       0.310       0.892
PClass[T.3rd]   -0.6282      0.132     -4.754      0.000      -0.887      -0.369
Age             -0.0392      0.008     -5.144      0.000      -0.054      -0.024
==============================================================================
"""
```

The `family = sm.families.Binomial()` tells us that we're fitting a logistic regression, since we are stating that the outcomes are from a Binomial distribution. (See the API documentation for a list of available distributions for GLMs).

The coefficients in a logistic regression are the *log-odds ratios*. To get the odds ratios, we would need to exponentiate them.

```
np.exp(mod_logistic.params.drop('Intercept'))
```

```
Sex[T.male]      0.071981
PClass[T.1st]    6.640989
PClass[T.2nd]    1.824486
PClass[T.3rd]    0.533574
Age              0.961581
dtype: float64
```

The intercept term in a logistic regression is **not** a log-odds ratio, so we omit it by using the `drop` function.

6.5.3 Survival analysis

Survival analysis or reliability analysis deals typically with data on time to an event, where this time can be *censored* at the end of observation. Examples include time to death for cancer patients, time to failure of a car transmission, etc. Censoring would mean that the subject is still alive/working when we last observed.

A common regression method for survival data is Cox proportional hazards regression. As an example, we will use a data set from a VA lung cancer study.

```
veteran = sm.datasets.get_rdataset('veteran', 'survival').data

mod_cph = smf.phreg('time ~ C(trt) + celltype + age + C(prior)',
  data = veteran, status = veteran.status).fit()
mod_cph.summary()
```

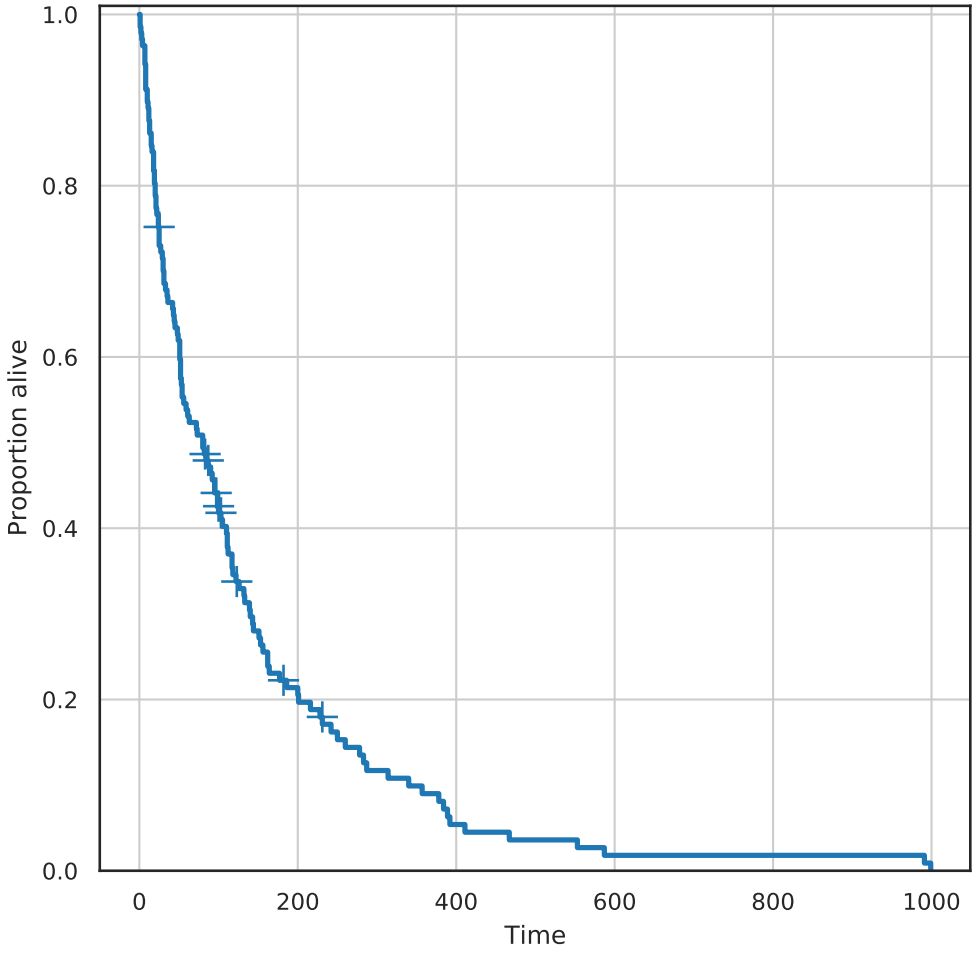
```
<class 'statsmodels.iolib.summary2.Summary'>
"""
                            Results: PHReg
=============================================================================
Model:                        PH Reg            Sample size:           137
Dependent variable:           time              Num. events:           128
Ties:                         Breslow
-----------------------------------------------------------------------------
                      log HR log HR SE   HR      t    P>|t|  [0.025 0.975]
-----------------------------------------------------------------------------
C(trt)[T.2]            0.1734   0.2016 1.1893  0.8600 0.3898 0.8011 1.7655
celltype[T.large]     -0.8817   0.2962 0.4141 -2.9761 0.0029 0.2317 0.7400
celltype[T.smallcell] -0.0956   0.2649 0.9088 -0.3609 0.7182 0.5407 1.5275
celltype[T.squamous]  -1.1738   0.2997 0.3092 -3.9173 0.0001 0.1718 0.5563
C(prior)[T.10]         0.0378   0.2064 1.0385  0.1833 0.8546 0.6930 1.5563
age                    0.0042   0.0096 1.0042  0.4401 0.6598 0.9855 1.0233
=============================================================================
Confidence intervals are for the hazard ratios
"""
```

For survival regression, we need to input the status of the subject at time of last follow-up, coded as 1 for failure/death, 0 for censored.

Question: Why did I use `C(trt)` instead of `trt` in the formula?

We can do a few more basic things for this data. First, let's draw the survival curve, which plots the proportion of subjects still alive against time, using the Kaplan-Meier method.

```
sf = sm.duration.SurvfuncRight(veteran.time, veteran.status)
sf.plot();
plt.grid(True);
plt.xlabel('Time');
plt.ylabel('Proportion alive');
plt.show()
```

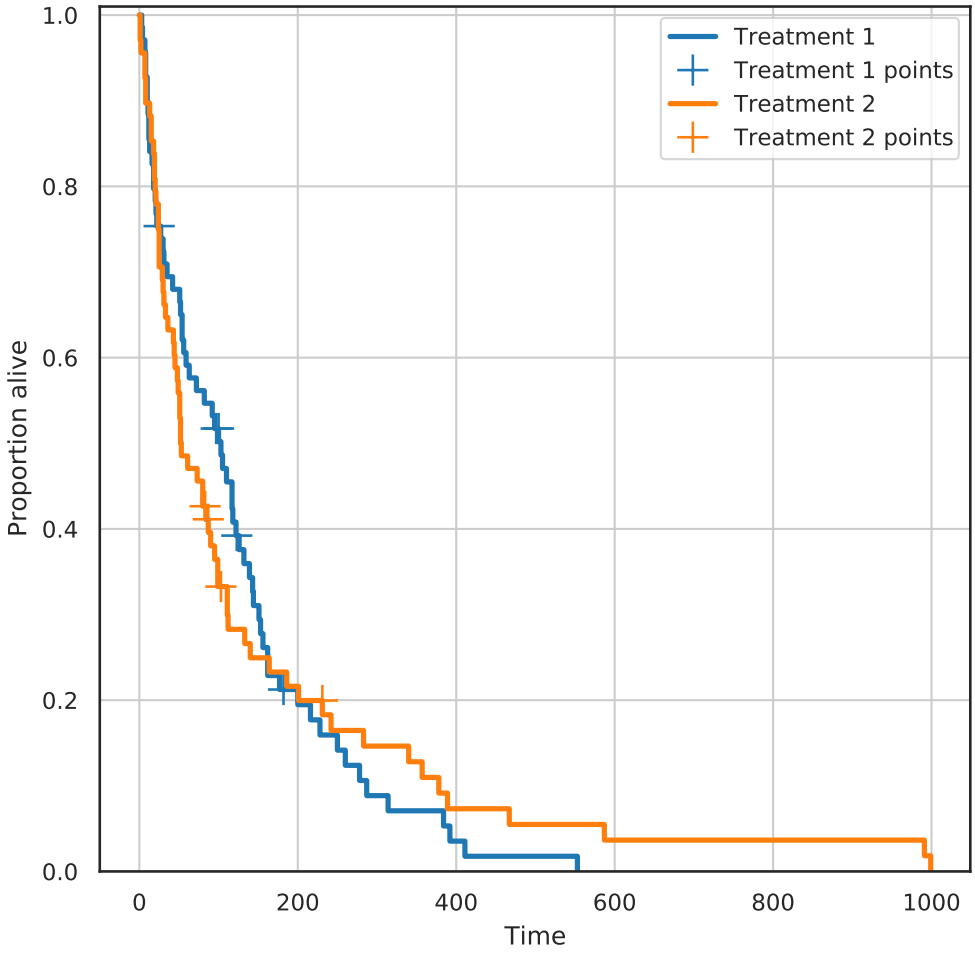


Suppose we now want to see if there is any difference between treatment groups.

```
sf1 = sm.duration.SurvfuncRight(veteran.time[veteran.trt==1], veteran.status[veteran.trt
sf2 = sm.duration.SurvfuncRight(veteran.time[veteran.trt==2], veteran.status[veteran.trt

fig, ax = plt.subplots()

plt.grid(True)
sf1.plot(ax); # Draw on previously defined axis
sf2.plot(ax);
```

```
<Figure size 600x600 with 1 Axes>
```

```
plt.xlabel('Time');
plt.ylabel('Proportion alive');
plt.legend(loc='upper right');
plt.show()
```



We could also perform a statistical test (the *log-rank test*) to see if there is a statistical difference between these two curves.

```
chisq, pval = sm.duration.survdiff(veteran.time, veteran.status, veteran.trt)
np.round(pval,3)
```

```
0.928
```

7 Machine Learning using Python

7.1 Scikit-learn

Scikit-learn (`sklearn`) is the main Python package for machine learning. It is a widely-used and well-regarded package. However, there are a couple of challenges to using it given the usual `pandas`-based data munging pipeline.

1. `sklearn` requires that all inputs be numeric, and in fact, `numpy` arrays.
2. `sklearn` requires that all categorical variables by replaced by 0/1 dummy variables
3. `sklearn` requires us to separate the predictors from the outcome. We need to have one `X` matrix for the predictors and one `y` vector for the outcome.

The big issue, of course, is the first point. Given we used `pandas` precisely because we wanted to be able to keep heterogenous data. We have to be able to convert non-numeric data to numeric. `pandas` does help us out with this problem.

1. First of all, we know that all `pandas` Series and DataFrame objects can be converted to `numpy` arrays using the `values` or `to_numpy` functions.
2. Second, we can easily extract a single variable from the data set using either the usual extracton methods or the `pop` function.
3. Third, `pandas` gives us a way to convert all categorical values to numeric dummy variables using the `get_dummies` function. This is actually a more desirable solution than what you will see in cyberspace, which is to use the `OneHotEncoder` function from `sklearn`.
    - This is generally fine since many machine learning models look for interactions internally and don't need them to be overtly specified. The main exceptions to this are linear and logistic regression. For those, we can use the formula methods described in the Statistical Modeling module to generate the appropriately transformed design matrix.
    - If the outcome variable is not numeric, we can `LabelEncoder` function from the `sklearn.preprocessing` submodule to convert it.

I just threw a bunch of jargon at you. Let's see what this means.

7.1.1 Transforming the outcome/target

```
import numpy as np
import pandas as pd
import sklearn
import statsmodels.api as sm
import matplotlib.pyplot as plt
```

```
import seaborn as sns

iris = sm.datasets.get_rdataset('iris').data
iris.head()
```

```
   Sepal.Length  Sepal.Width  Petal.Length  Petal.Width Species
0           5.1          3.5           1.4          0.2  setosa
1           4.9          3.0           1.4          0.2  setosa
2           4.7          3.2           1.3          0.2  setosa
3           4.6          3.1           1.5          0.2  setosa
4           5.0          3.6           1.4          0.2  setosa
```

Let's hit the first issue first. We need to separate out the outcome (the variable we want to predict) from the predictors (in this case the sepal and petal measurements).

```
y = iris['Species']
X = iris.drop('Species', axis = 1) # drops column, makes a copy
```

Another way to do this is

```
y = iris.pop('Species')
```

If you look at this, `iris` now only has 4 columns. So we could just use `iris` after the pop application, as the predictor set

We still have to update y to become numeric. This is where the `sklearn` functions start to be handy

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y = le.fit_transform(y)
y
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

Let's talk about this code, since it's very typical of the way the `sklearn` code works. First, we import a method (`LabelEncoder`) from the appropriate `sklearn` module. The second line, `le = LabelEncoder()` works to "turn on" the method. This is like taking a power tool off the shelf and plugging it in to a socket. It's now ready to work. The third line does the actual work. The `fit_transform` function transforms the data you input into it based on the method it is then attached to.

> Let's make a quick analogy. You can plug in both a power washer and a jackhammer to get them ready to go. You can then apply each of them to your driveway. They "transform" the driveway in different ways depending on which tool is used. The washer would "transform" the driveway by cleaning it, while the jackhammer would transform the driveway by breaking it.

There's an interesting invisible quirk to the code, though. The object `le` also got transformed during this process. There were pieces added to it during the `fit_transform` process.

```
le = LabelEncoder()
d1 = dir(le)
```

```
y = le.fit_transform( pd.read_csv('data/iris.csv')['species'])
d2 = dir(le)
set(d2).difference(set(d1)) # set of things in d2 but not in d1
```

```
{'classes_'}
```

So we see that there is a new component added, called `classes_`.

```
le.classes_
```

```
array(['setosa', 'versicolor', 'virginica'], dtype=object)
```

So the original labels aren't destroyed; they are being stored. This can be useful.

```
le.inverse_transform([0,1,1,2,0])
```

```
array(['setosa', 'versicolor', 'versicolor', 'virginica', 'setosa'],
      dtype=object)
```

So we can transform back from the numeric to the labels. Keep this in hand, since it will prove useful after we have done some predictions using a ML model, which will give numeric predictions.

7.1.2 Transforming the predictors

Let's look at a second example. The `diamonds` dataset has several categorical variables that would need to be transformed.

```
diamonds = pd.read_csv('data/diamonds.csv.gz')

y = diamonds.pop('price').values
X = pd.get_dummies(diamonds)

# Alternatively
# import patsy
# f = '~ np.log(carat) +  + clarity + depth + cut * color'
# X = patsy.dmatrix(f, data=diamonds)
```

```
type(X)
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
X.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53940 entries, 0 to 53939
Data columns (total 26 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   carat           53940 non-null  float64
 1   depth           53940 non-null  float64
 2   table           53940 non-null  float64
 3   x               53940 non-null  float64
 4   y               53940 non-null  float64
 5   z               53940 non-null  float64
 6   cut_Fair        53940 non-null  uint8  
 7   cut_Good        53940 non-null  uint8  
 8   cut_Ideal       53940 non-null  uint8  
 9   cut_Premium     53940 non-null  uint8  
 10  cut_Very Good   53940 non-null  uint8  
 11  color_D         53940 non-null  uint8  
 12  color_E         53940 non-null  uint8  
 13  color_F         53940 non-null  uint8  
 14  color_G         53940 non-null  uint8  
 15  color_H         53940 non-null  uint8  
 16  color_I         53940 non-null  uint8  
 17  color_J         53940 non-null  uint8  
 18  clarity_I1      53940 non-null  uint8  
 19  clarity_IF      53940 non-null  uint8  
 20  clarity_SI1     53940 non-null  uint8  
 21  clarity_SI2     53940 non-null  uint8  
 22  clarity_VS1     53940 non-null  uint8  
 23  clarity_VS2     53940 non-null  uint8  
 24  clarity_VVS1    53940 non-null  uint8  
```

```
 25  clarity_VVS2   53940 non-null  uint8
dtypes: float64(6), uint8(20)
memory usage: 3.5 MB
```

So everything is now numeric!!. Let's take a peek inside.

```
X.columns
```

```
Index(['carat', 'depth', 'table', 'x', 'y', 'z', 'cut_Fair', 'cut_Good',
       'cut_Ideal', 'cut_Premium', 'cut_Very Good', 'color_D', 'color_E',
       'color_F', 'color_G', 'color_H', 'color_I', 'color_J', 'clarity_I1',
       'clarity_IF', 'clarity_SI1', 'clarity_SI2', 'clarity_VS1',
       'clarity_VS2', 'clarity_VVS1', 'clarity_VVS2'],
      dtype='object')
```

So, it looks like the continuous variables remain intact, but the categorical variables got exploded out. Each variable name has a level with it, which represents the particular level it is representing. Each of these variables, called dummy variables, are numerical 0/1 variables. For example, `color_F` is 1 for those diamonds which have color F, and 0 otherwise.

```
pd.crosstab(X['color_F'], diamonds['color'])
```

```
color       D     E     F      G     H     I     J
color_F
0        6775  9797     0  11292  8304  5422  2808
1           0     0  9542      0     0     0     0
```

7.2 The methods

We will first look at supervised learning methods.

ML method	Code to call it
Decision Tree	`sklearn.tree.DecisionTreeClassifier`, `sklearn.tree.DecisionTreeRegressor`
Random Forest	`sklearn.ensemble.RandomForestClassifier`, `sklearn.ensemble.RandomForestRegressor`
Linear Regression	`sklearn.linear_model.LinearRegression`
Logistic Regression	`sklearn.linear_model.LogisticRegression`
Support Vector Machines	`sklearn.svm.LinearSVC`, `sklearn.svm.LinearSVR`

The general method that the code will follow is :

```
from sklearn.... import Machine
machine = Machine(*parameters*)
machine.fit(X, y)
```

7.2.1 A quick example

```
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor

lm = LinearRegression()
dt = DecisionTreeRegressor()
```

Lets manufacture some data

```
x = np.linspace(0, 10, 200)
y = 2 + 3*x - 5*(x**2)
d = pd.DataFrame({'x': x})

lm.fit(d,y)
```

```
LinearRegression()
```

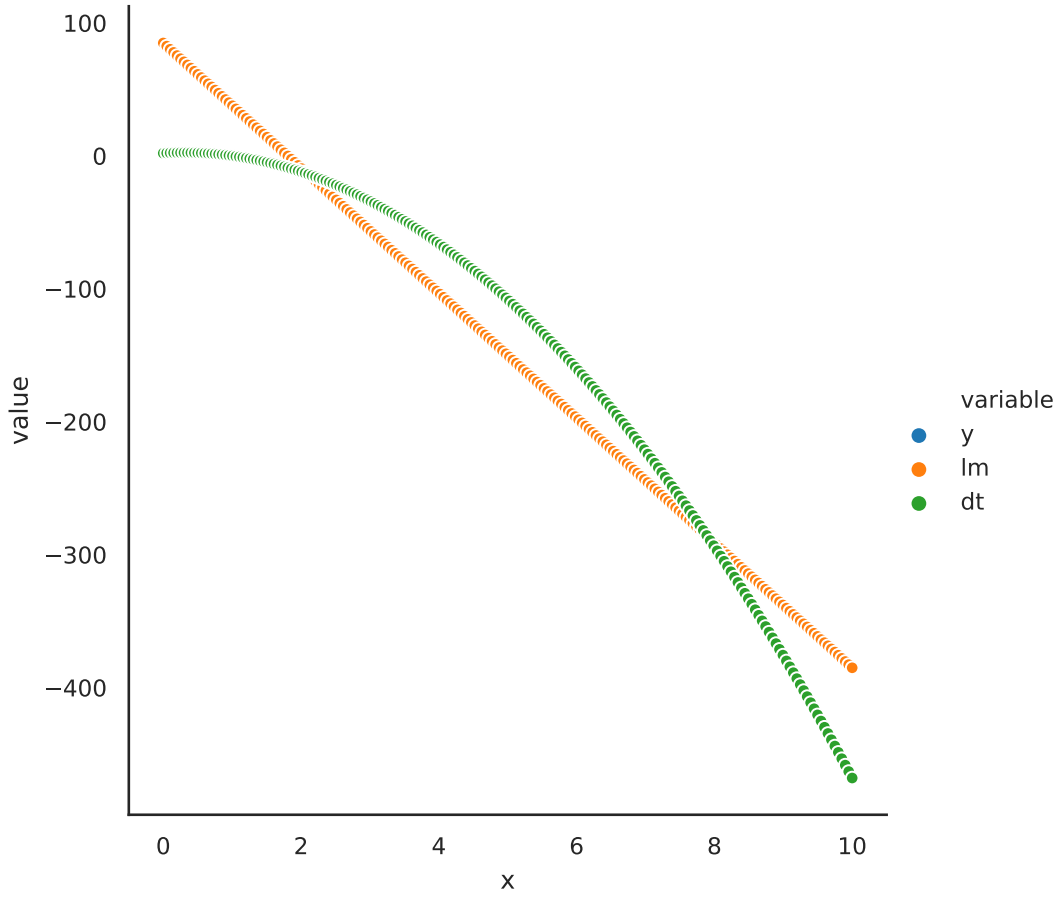
```
dt.fit(d, y)
```

```
DecisionTreeRegressor()
```

```
p1 = lm.predict(d)
p2 = dt.predict(d)

d['y'] = y
d['lm'] = p1
d['dt'] = p2

D = pd.melt(d, id_vars = 'x')

sns.relplot(data=D, x = 'x', y = 'value', hue = 'variable')
```

```
<seaborn.axisgrid.FacetGrid object at 0x127d1c760>
```

```
plt.show()
```



7.3 A data analytic example

```
diamonds = pd.read_csv('data/diamonds.csv.gz')
diamonds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53940 entries, 0 to 53939
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   carat    53940 non-null  float64
 1   cut      53940 non-null  object 
 2   color    53940 non-null  object 
 3   clarity  53940 non-null  object 
 4   depth    53940 non-null  float64
 5   table    53940 non-null  float64
 6   price    53940 non-null  int64  
```

```
 7   x         53940 non-null  float64
 8   y         53940 non-null  float64
 9   z         53940 non-null  float64
dtypes: float64(6), int64(1), object(3)
memory usage: 4.1+ MB
```

First, lets separate out the outcome (price) and the predictors

```
y = diamonds.pop('price')
```

For many machine learning problems, it is useful to scale the numeric predictors so that they have mean 0 and variance 1. First we need to separate out the categorical and numeric variables

```
d1 = diamonds.select_dtypes(include = 'number')
d2 = diamonds.select_dtypes(exclude = 'number')
```

Now let's scale the columns of `d1`

```
from sklearn.preprocessing import scale

bl = scale(d1)
bl
```

```
array([[-1.19816781, -0.17409151, -1.09967199, -1.58783745, -1.53619556,
        -1.57112919],
       [-1.24036129, -1.36073849,  1.58552871, -1.64132529, -1.65877419,
        -1.74117497],
       [-1.19816781, -3.38501862,  3.37566251, -1.49869105, -1.45739502,
        -1.74117497],
       ...,
       [-0.20662095,  0.73334442,  1.13799526, -0.06343409, -0.04774083,
         0.03013526],
       [ 0.13092691, -0.52310533,  0.24292836,  0.37338325,  0.33750627,
         0.28520393],
       [-0.10113725,  0.31452784, -1.09967199,  0.08811478,  0.11861587,
         0.14349912]])
```

Woops!! We get a numpy array, not a `DataFrame`!!

```
bl = pd.DataFrame(scale(d1))
bl.columns = list(d1.columns)
d1 = bl
```

Now, let's recode the categorical variables into dummy variables.

```
d2 = pd.get_dummies(d2)
```

and put them back together

```
X = pd.concat([d1,d2], axis = 1)
```

Next we need to split the data into a training set and a test set. Usually we do this as an 80/20 split. The purpose of the test set is to see how well the model works on an "external" data set. We don't touch the test set until we're done with all our model building in the training set. We usually do the split using random numbers. We'll put 40,000 observations in the training set.

```
ind = list(X.index)
np.random.shuffle(ind)

X_train, y_train = X.loc[ind[:40000],:], y[ind[:40000]]
X_test, y_test = X.loc[ind[40000:],:], y[ind[40000:]]
```

There is another way to do this

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y , test_size = 0.2, random_state
```

Now we will fit our models to the training data. Let's use a decision tree model, a random forest model, and a linear regression.

```
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor

lm = LinearRegression()
dt = DecisionTreeRegressor()
rf = RandomForestRegressor()
```

Now we will use our training data to fit the models

```
lm.fit(X_train, y_train)
```

```
LinearRegression()
```

```
dt.fit(X_train, y_train)
```

```
DecisionTreeRegressor()
```

```
rf.fit(X_train, y_train)
```

```
RandomForestRegressor()
```

We now need to see how well the model fit the data. We'll use the R2 statistic to be our metric of choice to evaluate the model fit.

```
from sklearn.metrics import  r2_score

f"""
Linear regression: {r2_score(y_train, lm.predict(X_train))},
Decision tree: {r2_score(y_train, dt.predict(X_train))},
Random Forest: {r2_score(y_train, rf.predict(X_train))}
"""
```

```
'\nLinear regression: 0.9202636015648039, \nDecision tree: 0.9999965428396391,\nRandom For
```

This is pretty amazing. However, we know that if we try and predict using the same data we used to train the model, we get better than expected results. One way to get a better idea about the true performance of the model when we will try it on external data is to do cross-validation.

7.3.1 Visualizing a decision tree

scikit-learn provides a decent way of visualizing a decision tree using a program called *Graphviz*, which is a dedicated graph and network visualization program.

```
import graphviz
from sklearn import tree

dt = DecisionTreeRegressor(max_depth=3)
dt.fit(X_train, y_train)
```

```
DecisionTreeRegressor(max_depth=3)
```

```
dot_data = tree.export_graphviz(dt, out_file=None,
                                feature_names = X_train.columns,
                                filled=True, rounded=True)
graph = graphviz.Source(dot_data);
graph
```

```
<graphviz.files.Source object at 0x13c780b20>
```

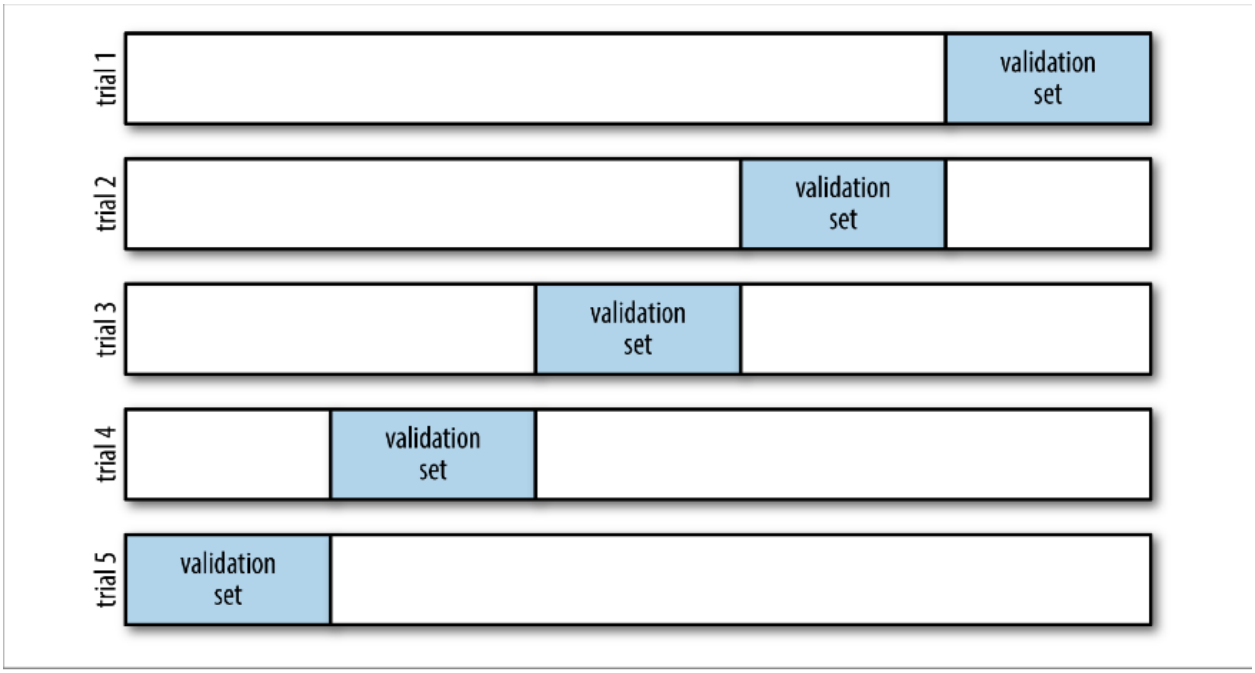
Alternatively,

```
tree.plot_tree(dt);
```

```
[Text(232.5, 404.25, 'X[0] <= 0.416\nmse = 15852573.11\nsamples = 43152\nvalue = 3925.87'),
0.183\nmse = 1250735.975\nsamples = 27947\nvalue = 1635.185'), Text(58.125, 173.25, 'X[4] <=
0.648\nmse = 267632.084\nsamples = 19909\nvalue = 1054.442'), Text(29.0625, 57.75, 'mse = 56
```

7.3.2 Cross-validation

In cross-validation, we split the dataset up into 5 equal parts randomly. We then train the model using 4 parts and predict the data on the 5th part. We do for all possible groups of 4 parts. We then consider the overall performance of prediction.



There is nothing special about the 5 splits. If you use 5 splits, it is called 5-fold cross-validation (CV), if you use 10 splits, it is 10-fold CV. If you use all but one subject as training data, and that one subject as test data, and cycle through all the subjects, that is called leave-one-out CV (LOOCV). All these methods are widely used, but 5- and 10-fold CV are often used as a balance between effectiveness and computational efficiency.

scikit-learn makes this pretty easy, using the `cross_val_score` function.

```
from sklearn.model_selection import cross_val_score
cv_score = cross_val_score(dt, X_train, y_train, cv=5, scoring='r2')
f"CV error = {np.round(np.mean(cv_score), 3)}"
```

```
'CV error = 0.874'
```

7.3.3 Improving models through cross-validation

The cross-validation error, as we've seen, gives us a better estimate of how well our model predicts on new data. We can use this to tune models by tweaking their parameters to get models that reasonably will perform better.

Each model that we fit has a set of parameters that govern how it proceeds to fit the data. These can bee seen using the `get_params` function.

```
dt.get_params()
```

```
{'ccp_alpha': 0.0, 'criterion': 'mse', 'max_depth': 3, 'max_features': None, 'max_leaf_node
```

```
le.get_params()
```

```
{}
```

> Linear regression is entirely determined by the functional form of the prediction equation,i.e., the "formula" we use. It doesn't have any parameters to tune per se. Improving a linear regression involves playing with the different predictors and transforming them to improve the predictions. This involve subjects called *regression diagnostics* and *feature engineering* that we will leave to Google for now.

We can tune different parameters for the decision tree to try and see if some combination of parameters can improve predictions. One way to do this, since we're using a computer, is a grid search. This means that we can set out sets of values of the parameters we want to tune, and the computer will go through every combination of those values to see how the model performs, and will provide the "best" model.

We would specify the values as a dictionary to the function `GridSearchCV`, which would optimize based on the cross-validation error.

```
from sklearn.model_selection import GridSearchCV
import numpy.random as rnd
rnd.RandomState(39358)
```

```
RandomState(MT19937) at 0x1399FDE40
```

```
param_grid = {'max_depth': [1,3,5,7, 10], 'min_samples_leaf': [1,5,10,20],
  'max_features' : ['auto','sqrt']}

clf = GridSearchCV(dt, param_grid, scoring = 'r2', cv = 5) # Tuning dt
clf.fit(X_train, y_train)
```

```
GridSearchCV(cv=5, estimator=DecisionTreeRegressor(max_depth=3),
             param_grid={'max_depth': [1, 3, 5, 7, 10],
                         'max_features': ['auto', 'sqrt'],
                         'min_samples_leaf': [1, 5, 10, 20]},
             scoring='r2')
```

```
clf.best_estimator_
```

```
DecisionTreeRegressor(max_depth=10, max_features='auto', min_samples_leaf=10)
```

```
print(clf.best_score_)
```

```
0.9646086949249106
```

So how does this do on the test set?

```
p = clf.best_estimator_.predict(X_test)
r2_score(y_test, p)
```

```
0.9657273806574082
```

So this predictor is doing slightly better on the test set than the training set. This is often an indicator that the model is overfitting on the data. This is probable here, given the extremely high R2 values for this model.

7.3.4 Feature selection

We can also use cross-validation to do recursive feature selection (or backwards elimination), based on a predictive score. This is different from usual stepwise selection methods which are based on a succession of hypothesis tests.

```
from sklearn.feature_selection import RFECV

selector = RFECV(lm, cv = 5, scoring = 'r2')
selector = selector.fit(X_train, y_train)
selector.support_
```

```
array([ True, False, False,  True, False, False,  True, False,  True,
        True,  True,  True,  True,  True,  True, False,  True,  True,
        True,  True, False,  True,  True,  True,  True,  True])
```

The support gives the set of predictors (True) that are finally selected.

```
X_train.columns[selector.support_]
```

```
Index(['carat', 'x', 'cut_Fair', 'cut_Ideal', 'cut_Premium', 'cut_Very Good',
       'color_D', 'color_E', 'color_F', 'color_G', 'color_I', 'color_J',
      'clarity_I1', 'clarity_IF', 'clarity_SI2', 'clarity_VS1', 'clarity_VS2',
       'clarity_VVS1', 'clarity_VVS2'],
      dtype='object')
```

This is indicating that the best predictive model for the linear regression includes carat, cut, color and clarity, and width of the stone.

7.4 Logistic regression

We noted that logistic regression is available both through **statsmodels** and through **scikit-learn**. Let's now try to fit a logistic regression model using **scikit-learn**. We will use the same Titanic dataset we used earlier.

```
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
from sklearn.linear_model import LogisticRegression

titanic = sm.datasets.get_rdataset('Titanic','Stat2Data').data.dropna()
titanic.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 756 entries, 0 to 1312
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Name      756 non-null    object
 1   PClass    756 non-null    object
 2   Age       756 non-null    float64
 3   Sex       756 non-null    object
 4   Survived  756 non-null    int64
 5   SexCode   756 non-null    int64
dtypes: float64(1), int64(2), object(3)
memory usage: 41.3+ KB
```

We will model `Survived` on the age, sex and passenger class of passengers.

```
from sklearn.model_selection import train_test_split

X = pd.get_dummies(titanic[['Age','Sex','PClass']], drop_first=True)
y = titanic.Survived

X_train, X_test, y_train, y_test = train_test_split(X, y , test_size = 0.2, random_state

lrm = LogisticRegression()
lrm.fit(X_train, y_train)
```

```
LogisticRegression()
```

There are a few differences that are now evident between this model and the model we fit using **statsmodels**. As a reminder, we fit this model again below.

```
titanic1 = titanic.loc[X_train.index,:]
titanic2 = titanic.loc[X_test.index,:]
mod_logistic = smf.glm('Survived ~ Age + Sex + PClass', data=titanic1,
  family = sm.families.Binomial()).fit()
mod_logistic.summary()
```

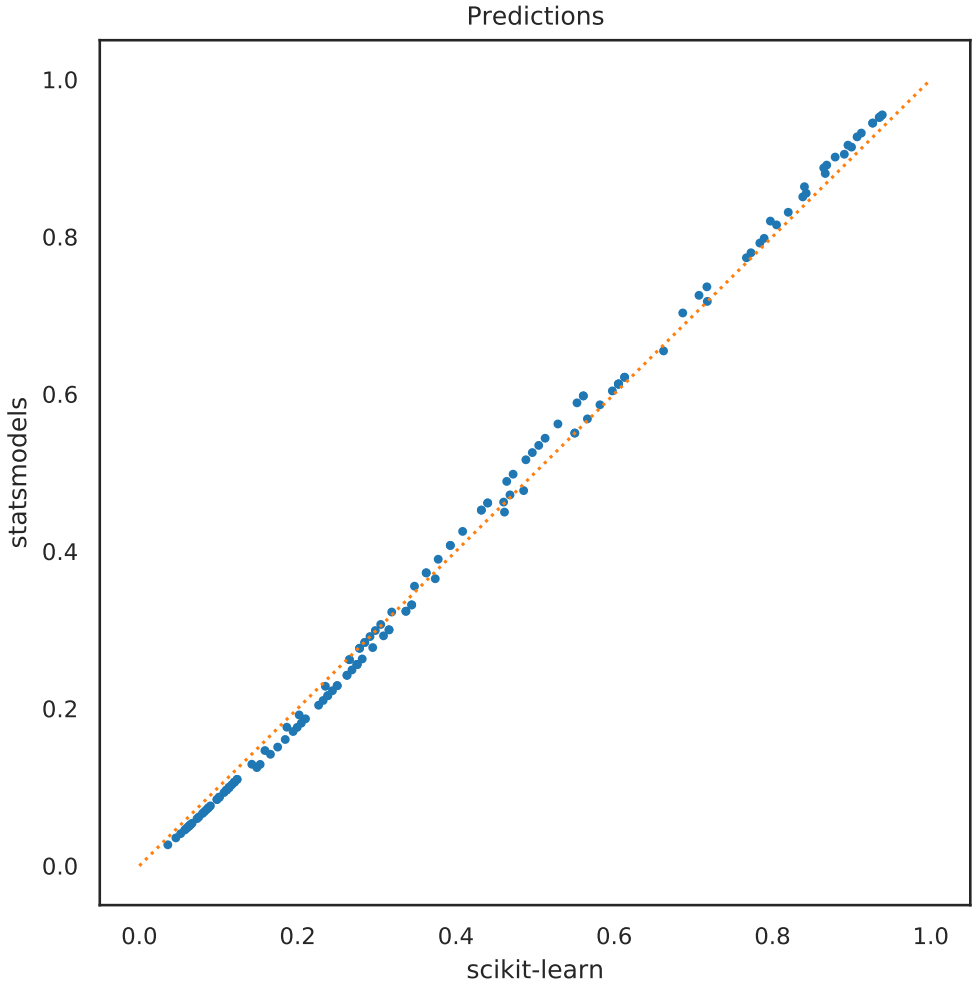
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                Survived   No. Observations:                  604
Model:                            GLM   Df Residuals:                      599
Model Family:                Binomial   Df Model:                            4
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -282.34
Date:                Thu, 04 Jun 2020   Deviance:                       564.68
Time:                        18:55:50   Pearson chi2:                     666.
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
-----
Intercept        3.6795      0.440      8.362      0.000       2.817       4.542
Sex[T.male]     -2.5138      0.221    -11.353      0.000      -2.948      -2.080
PClass[T.2nd]   -1.2057      0.290     -4.155      0.000      -1.774      -0.637
PClass[T.3rd]   -2.5974      0.305     -8.528      0.000      -3.194      -2.000
Age             -0.0367      0.008     -4.385      0.000      -0.053      -0.020
==============================================================================
"""
```

We can see the objects that are available to us from the two models using `dir(lrm)` and `dir(mod_logistic)`. We find that `lrm` does not give us any parameter estimates, p-values or summary methods. It is much leaner, and, in line with other machine learning models, emphasizes predictions. So if you want to find associations between predictors and outcome, you will have to use the **statsmodels** version.

Let's compare the predictions.

```
plt.clf()
p1 = lrm.predict_proba(X_test)[:,1]
p2 = mod_logistic.predict(titanic2)

plt.plot(p1, p2, '.');
plt.plot([0,1],[0,1], ':');
plt.xlabel('scikit-learn');
plt.ylabel('statsmodels');
```

```
plt.title('Predictions');
plt.show()
```

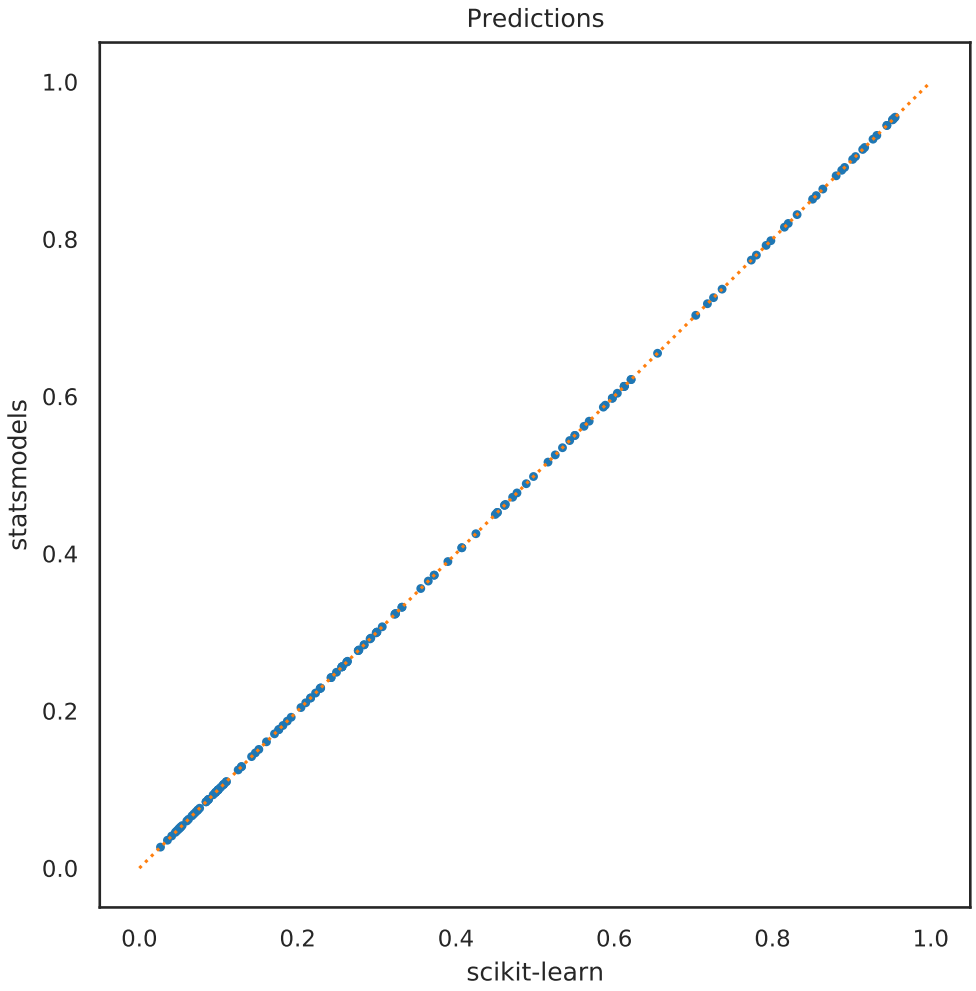


First note that the prediction functions work a bit differently. For `lrm` we have to explicitly ask for the probability predictions, whereas those are automatically provided for `mod_logistic`. We also find that the predictions aren't exactly the same. This is because `lrm`, by default, runs a penalized regression using the lasso criteria (L2 norm), rather than the non-penalized version that `mod_logistic` runs. We can specify no penalty for `lrm` and can see much closer agreement between the two models.

```
lrm = LogisticRegression(penalty='none')
lrm.fit(X_train, y_train)
```

```
LogisticRegression(penalty='none')
```

```
p1 = lrm.predict_proba(X_test)[:,1]

plt.clf()
plt.plot(p1, p2, '.');
plt.plot([0,1],[0,1], ':');
plt.xlabel('scikit-learn');
plt.ylabel('statsmodels');
plt.title('Predictions');
plt.show()
```



7.5 Unsupervised learning

Unsupervised learning is a class of machine learning methods where we are just trying to identify patterns in the data without any labels. This is in contrast to *supervised learning*, which are the modeling methods we have discussed above.

Most unsupervised learning methods fall broadly into a set of algorithms called *cluster analysis*. **scikit-learn** provides several clustering algorithms.

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters or distance threshold	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters or distance threshold, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
OPTICS	minimum cluster membership	Very large `n_samples`, large `n_clusters`	Non-flat geometry, uneven cluster sizes, variable cluster density	Distances between points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction.	Euclidean distance between points

We will demonstrate the two more popular choices – K-Means and Agglomerative clustering (also known as hierarchical clustering). We will use the classic Fisher's Iris data for this demonstration.

```
import statsmodels.api as sm
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.cluster import KMeans, AgglomerativeClustering

iris = sm.datasets.get_rdataset('iris').data
sns.relplot(data=iris, x = 'Sepal.Length',y = 'Sepal.Width', hue = 'Species');
```

```
<seaborn.axisgrid.FacetGrid object at 0x138c804c0>
```

The K-Means algorithm takes a pre-specified number of clusters as input, and then tries to find contiguous regions of the data to parse into clusters.

```
km = KMeans(n_clusters = 3)
km.fit(iris[['Sepal.Length','Sepal.Width']]);
```

```
KMeans(n_clusters=3)
```

```
km.labels_
```

```
array([2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1,
       1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0], dtype=int32)
```

```
iris['km_labels'] = km.labels_
iris['km_labels'] = iris.km_labels.astype('category')

sns.relplot(data=iris, x = 'Sepal.Length', y = 'Sepal.Width',
            hue = 'km_labels');
```

```
<seaborn.axisgrid.FacetGrid object at 0x1380bf490>
```

Agglomerative clustering takes a different approach. It starts by coalescing individual points successively, based on a distance metric and a principle for how to coalesce groups of points (called *linkage*). The number of clusters can then be determined either visually or via different cutoffs.

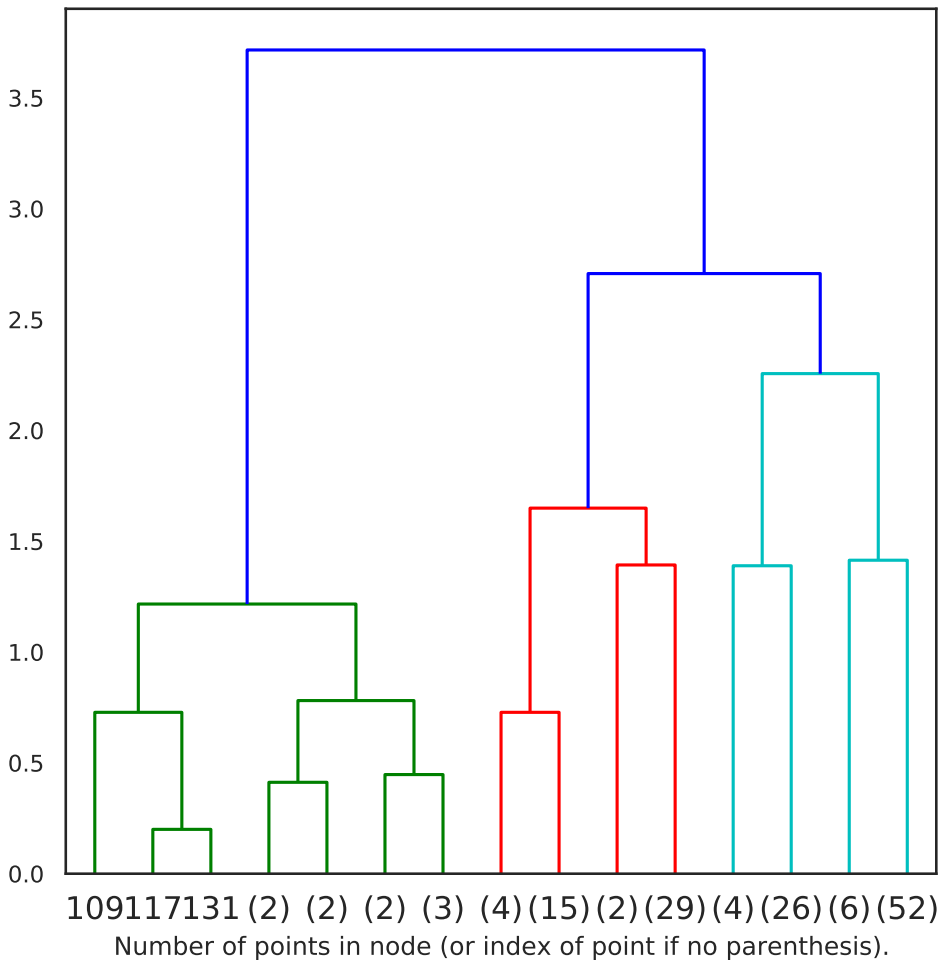
```
hc = AgglomerativeClustering(distance_threshold=0, n_clusters=None,
                             linkage='complete')

hc.fit(iris[['Sepal.Length','Sepal.Width']])
```

```
AgglomerativeClustering(distance_threshold=0, linkage='complete',
                        n_clusters=None)
```

```
hc.linkage
```

```
'complete'
```

```
from scipy.cluster.hierarchy import dendrogram

## The following is from https://scikit-learn.org/stable/auto_examples/cluster/plot_aggl
def plot_dendrogram(model, **kwargs):
    # Create linkage matrix and then plot the dendrogram

    # create the counts of samples under each node
    counts = np.zeros(model.children_.shape[0])
    n_samples = len(model.labels_)
    for i, merge in enumerate(model.children_):
        current_count = 0
        for child_idx in merge:
            if child_idx < n_samples:
                current_count += 1  # leaf node
            else:
                current_count += counts[child_idx - n_samples]
        counts[i] = current_count

    linkage_matrix = np.column_stack([model.children_, model.distances_,
                                      counts]).astype(float)

    # Plot the corresponding dendrogram
    dendrogram(linkage_matrix, **kwargs)

plot_dendrogram(hc, truncate_mode='level', p=3)
plt.xlabel("Number of points in node (or index of point if no parenthesis).")
plt.show()
```




```
hc = AgglomerativeClustering( n_clusters=3,
                              linkage='average')

hc.fit(iris[['Sepal.Length','Sepal.Width']]);
```

```
AgglomerativeClustering(linkage='average', n_clusters=3)
```

```
hc.labels_
```

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 0, 2, 0, 0, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 0, 2, 2, 2, 2, 0, 2, 2, 2, 1, 2, 2, 1, 0, 1, 2, 1,
```

```
       2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 1, 1, 1,
       2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2])
```

```
iris['hc_labels'] = pd.Series(hc.labels_).astype('category')

sns.relplot(data=iris, x = 'Sepal.Length', y= 'Sepal.Width',
            hue = 'hc_labels');
```

```
<seaborn.axisgrid.FacetGrid object at 0x13801bee0>
```

Play around with different linkage methods to see how these clusters change.

8 String manipulation

String manipulation is one of Python's strong suites. It comes built in with methods for strings, and the `re` module (for *regular expressions*) ups that power many fold.

Strings are objects that we typically see in quotes. We can also check if a variable is a string.

```
a = 'Les Miserable'

type(a)
```

```
<class 'str'>
```

Strings are a little funny. They look like they are one thing, but they can act like lists. In some sense they are really a container of characters. So we can have

```
len(a)
```

```
13
```

```
a[:4]
```

```
'Les '
```

```
a[3:6]
```

```
' Mi'
```

The rules are basically the same as lists. To make this explicit, let's consider the word 'bare'. In terms of positions, we can write this out.

index	0	1	2	3
string	b	a	r	e
neg index	-4	-3	-2	-1

We can also slices strings (and lists for that matter) in intervals. So, going back to `a`,

```
a[::2]
```

```
'LsMsrbe'
```

slices every other character.

Strings come with several methods to manipulate them natively.

```
'White Knight'.capitalize()
```

```
'White knight'
```

```
"It's just a flesh wound".count('u')
```

```
2
```

```
'Almond'.endswith('nd')
```

```
True
```

```
'White Knight'.lower()
```

```
'white knight'
```

```
'White Knight'.upper()
```

```
'WHITE KNIGHT'
```

```
'flesh wound'.replace('flesh','bullet')
```

```
'bullet wound'
```

```
' This is my song    '.strip()
```

```
'This is my song'
```

```
'Hello, hello, hello'.split(',')
```

```
['Hello', ' hello', ' hello']
```

One of the most powerful string methods is `join`. This allows us to take a list of characters, and then put them together using a particular separator.

```
' '.join(['This','is','my','song'])
```

```
'This is my song'
```

Also recall that we are allowed "string arithmetic".

```
'g' + 'a' + 'f' + 'f' + 'e'
```

```
'gaffe'
```

```
'a '*5
```

```
'a a a a a '
```

8.0.1 String formatting

In older code, you will see a formal format statement.

```
var = 'horse'
var2 = 'car'

s = 'Get off my {}!'

s.format(var)
```

```
'Get off my horse!'
```

```
s.format(var2)
```

```
'Get off my car!'
```

This is great for templates.

```
template_string = """
{country}, our native village
There was a {species} tree.
We used to sleep under it.
"""

print(template_string.format(country='India', species = 'banyan'))
```

```
India, our native village
There was a banyan tree.
We used to sleep under it.
```

```
print(template_string.format(country = 'Canada', species = 'maple'))
```

```
Canada, our native village
There was a maple tree.
We used to sleep under it.
```

In Python 3.6+, the concept of `f-strings` or formatted strings was introduced. They can be easier to read, faster and have better performance.

```
country = 'USA'
f"This is my {country}!"
```

```
'This is my USA!'
```

8.1 Regular expressions

Regular expressions are amazingly powerful tools for string search and manipulation. They are available in pretty much every computer language in some form or the other. I'll provide a short and far from comprehensive introduction here. The website regex101.com is a really good resource to learn and check your regular expressions.

8.1.1 Pattern matching

Syntax	Description
`.`	Matches any one character
`^`	Matches from the beginning of a string
`$`	Matches to the end of a string
`*`	Matches 0 or more repetitions of the previous character
`+`	Matches 1 or more repetitions of the previous character
`?`	Matches 0 or 1 repetitions of the previous character
`{m}`	Matches `m` repetitions of the previous character
`{m,n}`	Matches any number from `m` to `n` of the previous character
`\`	Escape character
`[ ]`	A set of characters (e.g. `[A-Z]` will match any capital letter)
`( )`	Matches the pattern exactly
`|`	OR

9 BioPython

BioPython is a package aimed at bioinformatics work. As with many Python packages, it is opinionated towards the needs of the developers, so might not meet everyone's needs.

You can install BioPython using `conda install biopython`.

We'll do a short example

```
from Bio.Seq import Seq

#create a sequence object
my_seq = Seq("CATGTAGACTAG")

#print out some details about it
print("seq %s is %i bases long" % (my_seq, len(my_seq)))
```

```
seq CATGTAGACTAG is 12 bases long
```

```
print("reverse complement is %s" % my_seq.reverse_complement())
```

```
reverse complement is CTAGTCTACATG
```

```
print("protein translation is %s" % my_seq.translate())
```

```
protein translation is HVD*
```

BioPython has capabilities for querying databases like `Entrez`, read sequences, do alignments using FASTA, and the like.