**Python**

**Functions** parentheses indicate that *print(‘hello world’)* is a function

**New line** \n

**Formatting**

In Python programming, code formatting is crucial. Things like indentation, spaces, tabulation will make a huge difference - and you will soon realise (even the hard way). So, if you're not using an editor of the one listed above, please make sure that your editor can properly setup spacing settings, removing any tab and use 4 spaces as indentation. This is not mandatory for the Python interpreter, but it is the general convention used.

In general, adding blank lines to appropriate places is a good idea. If you are asked to "fix code", feel free to add missing blank lines.

Capitalize names.capitalize() → input will begin from capital letter

Begin with capital letter names.title() → every word will begin from a capital letter

Caps lock names.upper()

Every word begins with a lower letter names.lower()

Like trim in Excel (delete excessive spaces) names.strip()

Underscore in numbers \_ → 100\_000\_000 is 100 000 000. Просто используется, чтобы визуально разбивать число для более понятного визуального восприятия. Число 1\_2\_3.45 = 123.45.

**Escape character \ (for example \"\" or \'\')**

Use backlash \ to get the double quotation marks around the *The Afterparty* value in the example below:

A screenshot of a computer

Description automatically generated with low confidence

**Cases**

**A screenshot of a computer

Description automatically generated with low confidence**

**Importing modules |** [**Python modules**](https://docs.python.org/3/py-modindex.html) **|**

**Libraries or packages |** [**Packages on Anaconda**](https://anaconda.cloud/package-categories)

[Math](https://docs.python.org/3/library/math.html) import math

[Statistics](https://docs.python.org/3/library/statistics.html#module-statistics) import statistics

[SQLite3](https://docs.python.org/3/library/sqlite3.html#module-sqlite3) import sqlite3

[Time](https://docs.python.org/3/library/time.html#module-time) import time

[Itertools](https://docs.python.org/3/library/itertools.html#module-itertools) import itertools

[Random](https://docs.python.org/3/library/random.html#module-random) import random

[Collections](https://docs.python.org/3/library/collections.html#module-collections) import collections

[Rich](https://github.com/Textualize/rich) Rich is library for rich text and beautiful formatting in the terminal pip install rich from richt import print

[chime](https://github.com/MaxHalford/chime) Sound notifications pip install chim import chime import chime.success() or chime.warning()

[schedule](https://github.com/dbader/schedule) pip install schedule import schedule

Python job scheduling for humans. Run Python functions (or any other callable) periodically using a friendly syntax.

[Faker](https://github.com/joke2k/faker) Generates fake data. pip install Faker import schedule

pandas ([1](https://github.com/pandas-dev/pandas)) import pandas as pd

functools import reduce

[tqdm](https://github.com/tqdm/tqdm) from tqdm import tqdm

warnings import warnings

**Import from a module:**

from math import sqrt as square\_root

square\_root(144)

>> 12.0

**Help / documentation**

**Use help() to get some information on the function:**

**A screenshot of a computer program

Description automatically generated with low confidenceA screenshot of a computer code

Description automatically generated with low confidence**

This is also true for your function/class:

**Use help() to get the notes / documentation of a class or a function:**

A screenshot of a computer program

Description automatically generated with medium confidence

**Variables**

Line of code. Like *a = 2* is a variable.

Variables can only contain letters, numbers, and underscores (symbol \_). Names can start only with a latter or an underscore, not with a number. Spaces are not allowed. Names should be descriptive without being too long: *mc\_wheels* is better than *wheels* or *number\_of\_wheels\_on\_a\_motorcycle*. Name should be the same. E.g. *name* = *name* but *name* != *Name*.

Python keywords that cannot be used as names:



Keywords don’t need parentheses like print().

Be careful about letter l (lowercase) and O (uppercase). They could be confused with 1 and 0.

**Type of data**

**Integer, int** numbers like 2, 3, 40

**Floating-point, float, число с плавающей запятой** numbers like 2.3, 4.32, 54.34

**String (строка букв), str** text like ‘Hello world’ or ‘2’ or ‘2.3’ (numbers in quotation marks)

Strings are contained either like this “double-quoted” or like this ‘single-quoted’.

Quotation inside a string:

Graphical user interface, text, application, email

Description automatically generated

Use triple-quote ‘’’ or “”” for preserving formatting between lines:

Graphical user interface, text, application, email, Teams

Description automatically generated

**Check type of data** type(‘hello world’) → str type(17) → int

Another type of data is None. It is Python’s ‘nothing’ value. It behaves just like any other value, and it's often used as a default value for different kinds of things.

Graphical user interface, text, application

Description automatically generated

**Booleans**

There are two Boolean values: True and False.

The *equal* symbol = is assigning a = 1 means *a* is set to *1*

Two *equal* symbols == are comparing a == 1 means *check if* a *is equal to* 1 / *does* a *equal* 1

In Python, a single equal sign = is used for assignment, while double equal signs == are used for comparison.

Alternative to *a == 1* is *(a == 1) == True*. However, because of readability simply use *a == 1*.

Empty or blank values are seen as False by Boolean and non-empty values are seen as True:

Graphical user interface, text, application, email

Description automatically generated

Background pattern

Description automatically generated with low confidence

**PEP 8:**

A screenshot of a computer code

Description automatically generated with low confidence

**Common Python idiom**

This is an example of a common Python idiom that leverages short-circuit evaluation and the truthiness of values:

True and 0 or "Name"

In Python, the and and or operators perform short-circuit evaluation. Here's how it works:

* For the and operator, if the left operand evaluates to False or a falsy value, the expression returns the left operand. Otherwise, it evaluates and returns the right operand. **In my words: it returns a truthy value.**
* For the or operator, if the left operand evaluates to True or a truthy value, the expression returns the left operand. Otherwise, it evaluates and returns the right operand. **It my words: it returns a truthy value.**

In the given code, the expression True and 0 or "Name" is evaluated as follows:

1. True and 0 is evaluated first. Since the left operand (True) is a truthy value, the expression proceeds to evaluate and return the right operand (0).
2. 0 or "Name" is evaluated next. Since the left operand (0) is a falsy value, the expression proceeds to evaluate and return the right operand ("Name").

Therefore, the result of the expression is "Name".

**Examples of data type:**

**LIST = [‘A’, ‘B’]**

A **list (square brackets)** is a collection of items that is stored in a variable (lists are collections of objects). The items should be related in some way, but there are no restrictions on what can be stored in a list.

Naming convention: plural name like *cars*, *dogs*. Each item is then a *car*, a *dog*. This gives you a straightforward way to refer to the entire list (dogs), and to a single item in the list (dog).

Example:

Graphical user interface

Description automatically generated with medium confidence

**Example of a list with two coupled variables like Name/Surname** people\_list = [("Alex", "De Jong"), ("John", "van Maarten"), ("Louis", "Verbeek")]

Retrieve a value: people[0] → ‘Alice’

Retrieve the last value: people[-1] → ‘Charlie’

Length of a list: len(people) → 3

Update list: people[1] = “John” → people → [“Alice”, “John”, “Charlie”]

Create an empty list: list = []

Convert a string (text) to a list:

 This method breaks a string into individual letters.

If you want to break a string into words use the split method:

A picture containing graphical user interface

Description automatically generated

To split a string by a delimiter add an optional argument:

Graphical user interface, text, application

Description automatically generated

**You can also join values in a list using a .join() and a delimiter. Two examples:**



The + operator concatenates lists:

Scatter chart

Description automatically generated with low confidence

The \* operator repeats the list a given number of times:

A screenshot of a computer

Description automatically generated with low confidence

Add 1 value to a list using .append() or add multiple values as 1 value:



If you want to add multiple values to the list you will need to use .extend():

A screenshot of a computer program

Description automatically generated with medium confidence

Delete the last 4 values in the list using del:



Remove one last item / remove one first item:



Create a new list with an added value:

A picture containing text, screenshot, font, white

Description automatically generated



A picture containing text, screenshot, font, line

Description automatically generated

**Create a copy of a list:**

In this example:

a = [1, 2, 3]

b = a

a change in a will lead to a change in b.

If you want to create b as a copy of a you shall use the .copy() code:

b = a.copy()

**Sort a list using .sort():**





**TUPLE = (‘A’, ‘B’)**

Graphical user interface, text, application

Description automatically generated

Tuples are basically lists that can never be changed.

Lists are quite dynamic; they can grow as you append and insert items, and they can shrink as you remove items. You can modify any element you want to in a list. Sometimes we like this behavior, but other times we may want to ensure that no user or no part of a program can change a list. That's what tuples are for.

Technically, lists are mutable objects and tuples are immutable objects. Mutable objects can change (think of mutations), and immutable objects can not change.

If you try to change the tuple using e.g. append() you will get an error. The same kind of thing happens when you try to remove something from a tuple, or modify one of its elements. In sum: once you define a tuple, you can be confident that its values will not change.

Tuple unpacking means:

A picture containing calendar

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

**Update a tuple:**

A screenshot of a computer

Description automatically generated with low confidence

**SET = {‘A’, ‘B’, ‘C’}**

A set object is an unordered collection of distinct hashable objects. Common uses include membership testing, removing duplicates from a sequence, and computing mathematical operations such as intersection, union, difference, and symmetric difference.

Sets are commonly used when you want to work **with unique elements** and **perform operations like finding intersections, unions, or differences** between sets. They **do not maintain order**, and **duplicate elements are automatically removed**.

**Create a set:**

Example 1: Example 2:

Graphical user interface, chat or text message

Description automatically generated with medium confidence A screenshot of a computer code

Description automatically generated with low confidence

**Difference between a *set* and a *dictionary*:**

A screenshot of a computer

Description automatically generated with medium confidence

**Create a set using set():**

set([‘apple’, ‘orange’])

**Operations in sets:**

**Example 1**

Graphical user interface, application, Word

Description automatically generated

**DIFFERENCE -** Find unique values in the first set that are not present in the second:

A screenshot of a computer

Description automatically generated with low confidence

**SYMMETRIC DIFFERENCE ^** Returns the elements that are unique to each set:

A screenshot of a computer

Description automatically generated with low confidence

**Example 2**

Text

Description automatically generated

**Example 3**

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated with medium confidence

**Add values to a set using .add():**

fruits = {'apple', 'banana', 'orange'}

fruits.add('mango')

**Remove values from a set using .remove() or .discard():**

Use .discard() when you want to *avoid an error* while removing a value that is (potentially) not present in your set.

fruits.remove(‘orange’)

fruits.discard(‘tomato’)

Number of values in a set using len():

len(fruits)

3

**Difference between lists, tuples, and sets:**

Lists:

* Ordered collection of elements (i.e. elements have a defined order and can be accessed by their index)
* Mutable (i.e. elements can be added, removed, and modified)
* Can contain duplicate elements
* Defined using square brackets [] or using the list() constructor function

Tuples:

* Ordered collection of elements
* Immutable (i.e. elements cannot be added, removed, or modified once the tuple is created)
* Can contain duplicate elements
* Defined using parentheses () or using the tuple() constructor function

Sets:

* Unordered collection of unique elements (i.e. no duplicates allowed)
* Mutable (i.e. elements can be added and removed)
* Elements are not accessed by index, but rather by value (i.e. you can check if an element is in a set, but you can't access it directly by index)
* Defined using curly braces {} or using the set() constructor function

Some similarities between lists, tuples, and sets:

* Can contain elements of any data type (strings, integers, floats, etc.)
* Can be iterated over using loops
* Can be used in combination with built-in Python functions like len() and sorted()

You would use a list when you need to store an ordered sequence of items that can be changed. For example, a list is useful when you want to keep track of a to-do list or a shopping list, where you may need to add or remove items.

You would use a tuple when you need to store an ordered sequence of items that cannot be changed. For example, a tuple is useful when you want to store the coordinates of a point in two-dimensional space, where the x and y values are fixed.

You would use a set when you need to store a collection of items that are unique and unordered. For example, a set is useful when you want to keep track of a list of unique words in a document, where the order of the words doesn't matter.

Here are some specific examples:

* List: todo\_list = ['shower', 'brush teeth', 'mow lawn', 'eat brains']
* Tuple: point = (3, 5)
* Set: unique\_words = {'apple', 'banana', 'orange', 'pear'}

Of course, these are just a few examples and there are many other use cases for lists, tuples, and sets depending on the specific problem you are trying to solve.

[] () {}

Table

Description automatically generated

[] () {}

Table

Description automatically generated

**Hashable / non-hashable objects**

In Python, a hashable object is an object that can be hashed, meaning it can be used as a key in a dictionary or an element in a set.

Hashable objects must have a hash value that remains the same throughout their lifetime (immutable). This means that if two hashable objects are equal, their hash values must be equal as well.

**Examples of hashable objects** include **integers, floats, strings, and tuples** (as long as their elements are also hashable).

On the other hand, non-hashable objects are objects that cannot be hashed, meaning they cannot be used as keys in a dictionary or elements in a set.

**Examples of non-hashable objects** include **lists, dictionaries, and other mutable objects**, which can change their contents or structure over time.

**DICTIONARY = {‘A’: 1, ‘B’: 2, ‘C’: 3}**

A dictionary is like a list, but more general.

In a list, the indices have to be integers *(numbers)*; in a dictionary they can be (almost) any type.



You can think of a dictionary as a **mapping** *(links)* between a set of indices (which are called keys) and a set of **values**. Each key maps to a value *(links to a value).*



The association of a key and a value is called a **key-value pair** or sometimes an **item**.

A dictionary connects two pieces of information. Those two pieces of information can be any kind of data structure in Python.

Create an empty dictionary: ruseng = {} or ruseng = dict()

What can be stored in a dictionary:

Text

Description automatically generated

**Merging two dictionaries using |**

Use | to make a **union of two dictionaries**. If there are some values that occur in both dictionaries, values from the second (last / right) dictionary will be taken in the merged dictionary:

A screenshot of a computer code

Description automatically generated with low confidence

**Difference between | (union operator) and |= (in-place union operator):**

A picture containing text, font, screenshot

Description automatically generated

| returns a new dictionary without modifying the original dictionaries: |= modifies the left-hand side dictionary in-place:

A screenshot of a computer code

Description automatically generated with low confidence A screenshot of a computer

Description automatically generated with low confidence

**.get()**

**Retrieve a value from a dictionary using .get().** It will retrieve a value if there is one and will return some text you write. It is useful while looking for values in dictionaries without getting the errors:

**Example 1** We have a *vocab* dictionary:

A picture containing text, font, white, algebra

Description automatically generated

It gets the value there is (*‘ni’*) and returns *‘Not yet known’* for a value that is absent:

A screenshot of a computer

Description automatically generated with medium confidence

**Example 2** The second part of the function (in the example above “Not yet known”) can be more than simply text. You can put there code that can be executed in case the first part of the function is not met:

A screenshot of a computer code

Description automatically generated with low confidence

**Example 3** .get() can be within .get(). In the example below a value is being looked in the second dictionary, that is an inverse of the first dictionary:

A screenshot of a computer code

Description automatically generated with low confidence

**Add new key-value pair**

Dictionary\_name[‘new key name’] = new\_value

Graphical user interface, text, email

Description automatically generated

Add a new key-value pair where the value is equal to a different key:

Text

Description automatically generated

**Modify a value**

Table

Description automatically generated with medium confidence

**Remove key-value pair using .pop()**

**Update: .pop() is used to remove the latest item together with selecting it (and not only in dictionaries but also in e.g. lists).**

If you give a name to the .pop line you can select the removed value.

Text

Description automatically generated

Remove key-value pair using .popitem(). It will remove the last inserted key-value pair & also return it.

Graphical user interface, text, application, email

Description automatically generated

**Lists in a dictionary**

**Example 1:**

A picture containing calendar

Description automatically generated

**Example 2:**

A screenshot of a computer code

Description automatically generated with medium confidence

Accessing values in lists:

A picture containing text, font, screenshot, white

Description automatically generated

**Traversing through dictionaries**

Text, letter

Description automatically generated

**Loop through all key-value pairs** using .items(). This code pulls all key-value pairs from a dictionary into a list of tuples.

Text

Description automatically generated

**Example 2 without and with list comprehensions:**



*Note that in both cases we are making an* ***inversed-dictionary****.*

Without: With:

A screenshot of a computer code

Description automatically generated with low confidence A screenshot of a computer code

Description automatically generated with low confidence

**Example 3 with list comprehensions:**



A screenshot of a computer

Description automatically generated with medium confidence

**Looping through all keys** in a dictionary using .keys() or without it:

A picture containing table

Description automatically generated Text

Description automatically generated

Both get the same result. This is because when you iterate over a dictionary using a for loop, it only iterates over the keys by default. In other words, in the line for key, value in my\_dict:, the variable key will be assigned to the key of each key-value pair in my\_dict, while the variable value will be left unassigned. To fix this error, you need to use the .items() method to iterate over both keys and values of the dictionary.

**Get the value of a specific key:**

Graphical user interface, text, application, email

Description automatically generated

Alternative example:

A picture containing diagram

Description automatically generated

Text

Description automatically generated

**Looping through all values** in a dictionary using .values():

Graphical user interface, text, application

Description automatically generated

**Traverse through two lists / iterate through two lists** at the same time using zip():



Enumerate through a dictionary / print a dictionary:



Enumerate through a dictionary with multiple values per key:

Graphical user interface, text, application, email

Description automatically generated

**IMPORT statistics**

**Calculate average of a dictionary’s values** using .mean():

Text

Description automatically generated

**Dictionary within a dictionary**

Print-out a dictionary within a dictionary:

Text

Description automatically generated

**Note on nesting**

While one level of nesting is really useful, nesting much deeper than that gets really complicated, really quickly. There are other structures such as classes which can be even more useful for modeling information. In addition to this, we can use Python to store information in a database, which is the proper tool for storing deeply nested information.

Often times when you are storing information in a database you will pull a small set of that information out and put it into a dictionary, or a slightly nested structure, and then work with it. But you will rarely, if ever, work with Python data structures nested more than one level deep

**ARRAYS**

In Python, an array is a collection of elements of the same data type, which are stored in contiguous memory locations. Unlike lists, arrays are designed for numerical data and provide more efficient storage and faster access to the elements.

Python provides the array module that allows you to create arrays. To create an array, you need to specify the data type of the elements, and then you can initialize the array with a sequence of values.

Here's an example:

A screenshot of a computer program

Description automatically generated with medium confidence

In this example, we create an array of integers using the array module with the type code 'i'. The type code is used to specify the data type of the array elements (in this case, integers). We also initialize the array with a list of values.

We can then access the elements of the array using indexing, change the value of an element using assignment, and iterate over the elements using a for loop.

**CLASSES**

**Object-oriented programming,** or **OOP** for short, focuses on building reusable blocks of code called **classes**. When you want to use a class in one of your programs, you make an object from that class, which is where the phrase "object-oriented" comes from. Python itself is not tied to object-oriented programming, but you will be using objects in most or all of your Python projects.

A **class** is a body of code that defines the attributes and behaviors required to accurately model something you need for your program. You can model something from the real world, such as a rocket ship or a guitar string, or you can model something from a virtual world such as a rocket in a game, or a set of physical laws for a game engine.

An **attribute** is a piece of information. In code, an attribute is just a variable that is part of a class.

A **behavior** is an action that is defined within a class. These are made up of **methods**, which are just functions that are defined for the class.

An **object** is a particular instance of a class. An object has a certain set of values for all of the attributes (variables) in the class. You can have as many objects as you want for any one class.

Classes are a way of combining information and behavior.

Example of a class:

A screenshot of a computer code

Description automatically generated with low confidence

Same with notes:

A picture containing text, screenshot, font

Description automatically generated

For example, let's consider what you'd need to do if you were creating a rocket ship in a game, or in a physics simulation. One of the first things you'd want to track are the x and y coordinates of the rocket. Here is what a simple rocket ship class looks like in code:

A picture containing text, screenshot, font

Description automatically generated

**FrozenSets**



**Named Tuples**

**COLLECTIONS** First you need to *import collections* 

Example 1:

Graphical user interface, text, application, email

Description automatically generated

Example 2:

Text

Description automatically generated

My example:



**Deques**

See also *Remove one last item / remove one first item:*

If you want to **add/remove a value** to a list *deques* can be an alternative.

**COLLECTIONS** First you need to *import collections* 

Then you need to transform your list into a *deque*:

Graphical user interface, text, application

Description automatically generated

Use *appendleft* or *append* to add values at the beginning/end:

Graphical user interface, text, application

Description automatically generated

If you want to select the first/last value use *popleft()* and *pop():*

Graphical user interface, application

Description automatically generated

These values will be removed from the *deque*. Use *list()* to transform the deque to a list.

**Operators and Operands**

Operators are special symbols that represent computations like addition and multiplication. The values the operator is applied to are called operands.

Operators:

+ addition

- subtraction

\* multiplication

/ division

\*\* exponentiation

% [modulo operator](https://www.freecodecamp.org/news/the-python-modulo-operator-what-does-the-symbol-mean-in-python-solved/) used to check if a number is чётный/нечётный (even/uneven number), или для того, чтобы узнать делится ли на какую-то цифру число (например *3* делится на *3*, *5* на *5*, *15* на *15* и так далее). Если делится, то результат будет *0*.

Text

Description automatically generated A picture containing text, font, symbol, logo

Description automatically generated

\*\* or *exponentiation* называется по-русски *в степени*. То есть 5 \*\* 2 это 5^2 или 52 или 5 \* 5.

В некоторых других языках символ ^ используется для *степени*, но в питоне это [bitwise operator](https://wiki.python.org/moin/BitwiseOperators) called XOR.

**Comparing operators**

Graphical user interface, text, application

Description automatically generated

There can also be multiple comparisons.

Graphical user interface, text

Description automatically generated

Examples:

Graphical user interface, text, application

Description automatically generated

**Order of operations**

A picture containing text

Description automatically generated

**String operations**

There are mostly two: *+* and *\**

Graphical user interface, text, application, email

Description automatically generated

More elaborate:

Graphical user interface, text, application, email

Description automatically generated

**Floating-point operations**

Could have unexpected results like here:

Graphical user interface, text, application

Description automatically generated

Don’t worry, this is the way computers represent numbers internally.

More elaborate:

Graphical user interface, text, application, email

Description automatically generated

**Commenting and comments**

Use hashtag symbol *#* to leave a comment.

Everything from the # to the end of the line is ignored.

Comments are most useful when they document non-obvious features of the code. It is reasonable to assume that the reader can figure out what the code does; it is much more useful to explain why.

Example:





Another way to leave a comment is using “””:

Text

Description automatically generated

**Functions**

An example of a function is *print()*. If you just execute *print* you will get then <function print> as output.

Functions do something when they are *called* (executed in my words) by typing their names and parentheses. Inside the parentheses, we can pass some arguments too. In *print(‘hello’)* the function is *print* and we give it one argument *‘hello’*.

Sometimes people think that doing func = print('hello') means that Python is going to print hello every time we type func. But this is not correct! print('hello') runs print right away, and if we type func later, it's not going to run print('hello') again.



In the context of programming, a function is a named sequence of statements that performs a computation.

When you define a function, you specify the name and the sequence of statements. Later, you can “call” (sometime "invoke") the function by name.

The left side of a function has always to be a name like a = 2 +2, not 2 + 2 = a.

**return statements in a function**

Be consistent in return statements. Either all return statements in a function should return an expression, or none of them should. If any return statement returns an expression, any return statements where no value is returned should explicitly state this as return None, and an explicit return statement should be present at the end of the function (if reachable):

Correct: Wrong:

A screenshot of a computer program

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidence

Я понимаю это так: если ты указываешь *ЕСЛИ должно произойти Х* (в данном случае *число должно быть больше или равно нулю*), то указывай также ЧТО должно произойти, если это не произойдёт. В данном случае или пиши else: return None или return None.

**Example of a function:**

A screenshot of a computer

Description automatically generated with medium confidence

dir() around a (for example) *list* gives an overview of the commands:

A picture containing text, document, screenshot, number

Description automatically generated

**Module**

Python has a *math* module that provides most of the familiar mathematical functions.

A *module* is a Python file that contains a collection of related functions.

To access a function in a module you need to specify the name of the module then dot (also known as a period) then the name of the function. Like *math.log10()*. This format is called dot notation (запись через точку).

**Function annotations**

You can attach metadate to the parameters of a function declaration and its return value.

Example:

A screenshot of a computer code

Description automatically generated with medium confidence

Use .\_\_annotations\_\_ to access function annotations:

A screenshot of a computer

Description automatically generated with low confidence

Use .\_\_doc\_\_ to see function comments:

A picture containing text, screenshot, font, line

Description automatically generated

* Each argument in the function declaration may have an annotation expression preceded by :.
* If there is a *default value*, the annotation goes between the argument name and the = sign.
* To annotate the return value, add -> and another expression between the ) and the : at the tail of the function declaration.

The expressions may be of any type. The most common types used in annotations are classes, like str or int, or strings, like 'int > 0'.

**Python functions**

**print()**

A picture containing text, font, screenshot, white

Description automatically generated

**input()**

Use input() to get start a input line to get some input from the user.

Example: After typing in my name:

A white box with red text

Description automatically generated with low confidence A picture containing text, font, screenshot, line

Description automatically generated

If you assign the input() function to a *name* like in the example above my\_name, you will get the *value* you got from the input while running the function:

A screenshot of a computer

Description automatically generated with low confidence

**round()**

A picture containing text, font, screenshot

Description automatically generated

**ord()**

Given a string representing one Unicode character, return an integer representing the Unicode code point of that character. For example, ord('a') returns the integer 97 and ord('€') (Euro sign) returns 8364. Inverse of chr().

**\_\_main\_\_ function**

Graphical user interface, text, application, email

Description automatically generated

Своими словами: ты написал 2 функции и ещё последнюю функцию назвал main, которая сводит первые две функции в одну. Если ты хочешь чтобы твоя третья функция main сразу же была активизирована, то ты пишешь в конце if \_\_name\_ и так далее. Пример из Chart GPT:



Альтернативой может быть просто main(), но тогда:

Text, letter

Description automatically generated

**from math module functions**

from math import sqrt

Math.log() compute logarithms

Math.sin() get the sinus, also .cos, .tan etc.

Math.pi gets the π value / pi value

Math.sqrt() get the square root of a number like math.sqrt(16) = 4.0

Math.exp ?

Math.hypot(x, y) Calculates the Euclidean norm (length of a vector with two components). Takes two numeric arguments and returns the square root of their sum of squares: x2 + y2.

**from random module functions**

from random import randint

randint() generate a random number between 0 and 10: randint(0, 10)

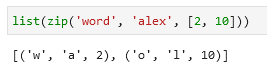
**zip()**

The zip() function takes multiple iterables as arguments and returns an iterator that generates tuples containing elements from each iterable. Each tuple in the resulting iterator represents a pair of corresponding elements from the input iterables.

In other words: it takes your input and creates a tuple of lists where the first list has only the first values, the second only the second values etc. It will be as long as the shortest value in your input.

Example 1: Example 2:

A picture containing text, font, screenshot, white

Description automatically generated 

**Higher-order functions**

Sort a list of words by its length (**sorted(),** **map()**):

A picture containing text, font, screenshot, algebra

Description automatically generated

**map() map(function, iterable)**

Iterate over an iterable (e.g. a list, tuple, string) using a function provided. In the example below iterate over list *numbers* using the *square* function:

A screenshot of a computer code

Description automatically generated with low confidence

**Note:** **map() returns an object and not a list**. That is why you need to **convert it to a list using the list()** function.

**for loop or map()**

The same is achieved by the following two examples below. However, the second example with the map() function is more concise and can be considered as a shorter and sometimes more efficient alternative to the first example with the for loop.

A screenshot of a computer code

Description automatically generated with low confidence

**lambda expression**

One way to write small functions is to use the lambda expression. lambda takes a number of parameters and an expression combining these parameters, and creates an anonymous function that returns the value of the expression:



A screenshot of a computer program

Description automatically generated with low confidence

From here <https://docs.python.org/3/howto/functional.html>

**Converting data / data conversion / datatype conversion / change data type**

Change data to integer (to int) int(‘32’) → 32

won’t work on *int(‘dadw’)*

*int(2.99) → 2* it will drop the part after the point (here .99)

Convert data to float float(3) → 3.0 float(‘3.2342’) → 3.2342

Convert data to string (to str) str(32) → ‘32’

**If-statements**

There two are the same:

Graphical user interface

Description automatically generated with medium confidenceA picture containing diagram

Description automatically generated

However the second is better since by default python considers blank values as false. So we can say that if message is not blank then print what was entered, otherwise print that you did not enter anything. Check the Boolean section for more info.

Same code (last is the best):

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Another example:

Graphical user interface, text, application

Description automatically generatedText

Description automatically generated

Make a space between two values:



A statement should be equal to null (should be blank):

Graphical user interface, text, application, email

Description automatically generated

**Function definition / define a function**

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Function example:

Graphical user interface, text, application, email

Description automatically generated

Function example:



Function example:

Graphical user interface, text, application

Description automatically generated

**Setting a default value in a function.** In this case the default value is *‘everyone’:*

Graphical user interface, text, application, email

Description automatically generated

If you have multiple arguments like def thank\_you(name, surname) you don’t need to input your arguments like thank\_you(‘John’, ‘Smit’). This will also work: thank\_you(surname=’Smit’, name=’John’). By keywording arguments you can place them in arbitrary order.



Text

Description automatically generated with low confidence

**Assigning two or more names to a tuple / extracting two or more variables from a tuple:**

Graphical user interface, application

Description automatically generated

1. If we have a list of elements the first and last variables will always select first and last elements. E.g. if we have 10 elements and 2 variables we will only select first and last element.

2. If I add extra variables between these (first and last) variables I will only select 2nd, 3rd etc. elements, but I other elements (e.g. from 4 to 9) will not be selected.

3. If I want to assign all other elements e.g. from 2 to 9 to a variable I need to add an asterisk before that variable.

4. Asterisk should be before that variable without any spaces.

Example:



a = 1

c = [2, 3, 4]

b = 5

**VSCode executing code**



**Sequences**

* **Container sequences** can hold items of different types. They hold reference to the objects they contain, which may be of any type. Examples:
  + list
  + tuple
  + collections.deque
* **Flat sequences** hold items of one type. They physically store the value of each item within its own memory space, and not as distinct objects. They are more compact, but are limited to holding primitive values like characters, bytes, and numbers. Examples:
  + str
  + bytes
  + bytearray
  + memoryview
  + array.array

Another way of grouping sequence types is **mutability**:

* **Mutable** sequences:
  + list
  + bytearray
  + array.array
  + collections.deque
  + memoryview
* **Immutable** sequences:
  + tuple
  + str
  + bytes



**for loop / for iteration loop / iterations / traversing**

This most widely used iteration mechanism in Python. Every sequence can be iterated (element by element). Also while loops are permitted, but the *for* loop is the one you’ll see and use most of the time. Strings and lists are iterable, but integers and floats are not

Example:

A picture containing text

Description automatically generated

Printing out names from the list called *names:*





Example of a *for* loop that prints out people’s names using a function *greeter:*

A picture containing diagram

Description automatically generated

*for* \*arbitrary name of the variables in the list\* *in* \*name of the list\*:

\*function name\*(\*arbitrary name of the variables in the list\*)

Example of a for loop with two or more variables:



**break and continue statements / keywords in a for loop:**

The example below illustrates the break statement in an *infinite loop*. The loop is running until the input is either *yes* or *y*. Then the loop is stopped using the break statement at the end.

Example 1:

A screenshot of a computer code

Description automatically generated with low confidence

Other examples:



Graphical user interface, text

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

**Enumerate a loop:**

Instead of *for dog in dogs* you add *for index, dog* and you add *enumerate(dogs):*

Graphical user interface, text, application

Description automatically generated

The value in the variable index (i) is always an integer (number). If you want to print it in a string turn the integer into a string like here:



Another example for **enumerating in a pair**:

Text

Description automatically generated with low confidence

**Manually iterating**

Use .\_\_next\_\_() to iterate manually through a string:

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

Description automatically generated

Alternative function that does the same thing is next():

Graphical user interface, text, application

Description automatically generated

This examples uses enumerate() and next() to go through the given list (manually iterate through a list):

A screenshot of a computer program

Description automatically generated with low confidence

**Iter() iteration**

There's a built-in function called iter() that converts anything iterable to an iterator.



Iterating using iter(): Alternative:

Text

Description automatically generated Graphical user interface, text, application

Description automatically generated

**Retrieve a specific part of e.g. a list:**

Make use of an additional row captures only the needed rows. In this case it is the *first\_batch = usernames[0:3]* row that selects only the first 3 items in the list:

Graphical user interface, text

Description automatically generated

**Examples of iterators**

enumerate(iterable)

range(start, end, step)

dictionary.keys()

dictionary.values()

dictionary.items()

Example of a .range():  
Graphical user interface, application

Description automatically generated

**while loops**

while loops are used when you want to repeatedly execute a block of code as long as a certain condition is true.

Basic structure of a while loop:

A picture containing text, font, screenshot, white

Description automatically generated

The loop starts by evaluating the condition. If the condition is true, the code inside the loop is executed. Afterward, the condition is checked again, and if it is still true, the code is executed again. This process continues until the condition becomes false. Once the condition is false, the loop is exited, and the program continues with the next line of code after the loop.

**Example 1: Example 2:**

A screenshot of a computer

Description automatically generated with low confidence A screenshot of a computer program

Description automatically generated with low confidence

**Example 3** (empty list [], .randrange(), .append())**:**

A screenshot of a computer

Description automatically generated with medium confidence

**Functions**

range() range(start, end, step) Generates a list of numbers between *start* and *end* with a given *step:* **range(0, 20, 2)**

The last value will not be included.

Graphical user interface, text, application

Description automatically generated

**Check if data exists in your list/tuple/set:**

A picture containing calendar

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**IMPORT statistics**

Calculate average of a dictionary’s values using .mean():

Text

Description automatically generated

|  |  |
| --- | --- |
| **Easter egg** | Import this |
| **Import pandas** | import pandas as pd |
| **Import NumPy** | import numpy as np |
| **Series (a column)** | pd.Series([21, 22, 23, 24], name = ‘age’)  *age = column name* |
| **Select an element (row; from 0 to …)** | table\_name[3]  *Select the 4th row* |
| **Make a custom axis to a Series** |  |
|  |  |
| **Index** is also called an **axis**; each element is called axis label.  Data in columns is called as **values**.  A series ideally should have the same datatype throughout its values (same format for the whole column). |  |
| **Dataframe** | A series is a column and a dataframe has multiple columns |
| **Create a dataframe** |  |
| **Columns and rows in a dataframe (table)** | df.shape  (3, 7)  3 rows and 7 columns |
| **Axis 0**  **Axis 1** | Vertical axis (rows)  Horizontal axis (column names) |
| **Select a column** | df[‘ADDRESS’]  Returns a series |
| **Select several columns** | df[[‘ADDRESS’, ‘CITY’]] |
| **Retrieve axis 1 / column names / information** | df.columns  Index([‘first\_name’, ‘last\_name’, ‘email’], dtype=‘object’) |
| **Retrieve axis 0 / vertical axis information** | df.index  RangeIndex(start=0, stop=3, step=1) or min 0, max 3 (3 rows) |
| **Retrieve a row** |  |
| **Make a column as an index** | E.g. you want to filter by last name. Instead of the default 0, 1, 2 etc. you can make the index as Nield, Scala Morrison.  *df.set\_index(‘column\_name’, inplace = True*) inplace is true so it edits the existing df instead of creating a new one  *df.loc[‘Scala’]* you can now use the last name to search for this row  **df.reset\_index(inplace = True)** reset the axis to default |
| **Copy a dataframe** | df2 = df.copy() |
| **New line \n** | Print(‘line1 \nline2’) |
| **Tab \t** | Print(‘line1 \tline2’) |
| **Raw string** | If you don’t want Python to see *\n* as code but part of the text (part of the string) then add an *r* add the beginning of the code.  Print(r’line1 \nline2’)  Output will not be  line 1  line 2  But line1 \nline2 |
| **Check the index of a value** | df = [ 1, 2, 3, 4]  df.index(2)  1 |
|  |  |

**Importing Data**





[**Import CSV / read csv / load csv**](https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html)

Read csv as text:



Read csv into as a pandas dataframe:



**Read csv alternative:**

df = pd.read\_csv('https://raw.githubusercontent.com/thomasnield/machine-learning-demo-data/master/regression/winequality-red.csv')

df

**With *header* and *names*:**



**Alternative:**

Import pandas as pd

url = 'https://raw.githubusercontent.com/thomasnield/machine-learning-demo-data/master/timeseries/datetime\_formatting.csv'

df = pd.read\_csv(url)

df

**Format dates columns as dates while importing csv:**



**SQL** [**pd.read\_sql**](https://pandas.pydata.org/docs/reference/api/pandas.read_sql.html)

Import an SQL database:



Alternative:



If you don’t *parse\_dates*:

With parsing and without:



**JSON**

Read a JSON file:



Read a JSON file as a pandas dataframe:



**Selecting Rows and Columns**

**loc and iloc**

* **loc** works on **labels** assigned to the axis, **iloc** works on **numbers**.
* E.g. if axis is surname then **loc** will work only on the surname like ‘**vanli’** while **iloc** will only work on **numbers**.

**Select** (first two) **rows** df.iloc[0:2] or df.iloc[:2]

**Exclude first row** df.iloc[1:]

**Select all** df.iloc[:]

**All rows and columns 2-3** df.iloc[:, 1:3] you are selecting columns 2 and 3 with indexes 1 and 2

**Select last two columns** df.iloc[:, -2:] count from 0 to -1, -2 etc. from right to left. Select the column you want to have and with : you will select everything to the right

**Select the last row** df.iloc[-1]

**Select all rows and last column** df.iloc[:, -1]

**Select all rows and all columns except for the last column** df.iloc[:, :-1]

**Select all rows and reverse them** df2 = df1[::-1]

If you want to get the first 2 rows you need to select the third index, in this case 2:



**Select specific columns and rows**. In this case you will select column index 0 and column index 2 (first\_name and email). In the second query you will select second row and third column



**Select rows and columns using loc** df.loc[["samiam","thomasnield"], "email"]

**Reset index** df.reset\_index(inplace = True)

**Select values that start with a specific letter** condition = df["username"].str.startswith("s") username = column, s = letter start

df[condition]

**Multiple conditions. AND & OR are & and |** condition = df["username"].str.startswith("s") & df["email"].str.contains("gmail")

df[condition]

**Get gata in a column that equals to a value** df[df[‘column’].eq(‘value’)] *(filter data to a value)*

[**at and iat**](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iat.html)

Similar to loc and iloc there is also an *at* and *iat*. These return a single value at a specific row and column index using numeric or labelled indices respectively.

**Drop columns and rows**





**Remove columns:**



**Remove columns by selecting specific columns like column 1 and 4:**



**Adding rows and columns (appending) / Joining concatting**

**Add a column at the end of the dataframe:**



**Add a column at a specific place:**



**Add a row:**



**Using concat (merge two datasets):**



**Merge two datasets using .concat()** In the example below the first *df* is the first dataset and *df.loc[0:2,:]* is the second.

Graphical user interface, text, application, email

Description automatically generated

**Updating data**

**Making a column in caps lock (upper):**



**Updating on a condition:**



**Update row on condition:**



**Unpivoting data (melting):**



Id\_vars = untouched columns

Value\_vars = columns that will be unpivoted

Var\_name = column names will be moved to this column

Value\_name = column values will be moved to this column



**Sorting, Casting, and Categories**

**Datatypes:**



**Timestamp example (date)** pd.Timestamp(‘20230130’) → 2023-01-30

**Difference between days** pd.Timestamp('20230130') - pd.Timestamp('20230127') → 3 days

**View datatypes** df.dtypes



You will get column name on the left (float, int, datatime etc.) and datatype on the right.

**Change a column to a different datatype** df[‘columnname’] = df[‘columnaname’].astype(‘bool’)

**Sort by 2 columns (first lightning, then rain inches)** df.sort\_values(by=["lightning","rain\_inches"])

df.sort\_values(by=["lightning","rain\_inches"],ascending=[False,True])

df.sort\_values(axis=0, ascending = True) *ascending if one value*

When using the sort methods, remember to **add the inplace=True parameter if you want to replace the existing dataframe** with the sorted one.

Sort by an row index or columns index df.sort\_index(axis = 0) for rows

df.sort\_index(axis = 1) for columns

**Replace a column with a *category* datatype:**

cat\_type = pd.CategoricalDtype(categories=["CLEAR", "MINOR", "MAJOR", "SEVERE"], ordered=True)

df["severity"] = df["severity"].astype(cat\_type) severity = column you want to replace

df.sort\_values(by=["severity"])

If you apply a categorization on a column that has values not mapping to any category, then those will become NA values.

**Python if/elif/else category function:**



**Apply this function to the wind\_speed\_mph column:**



**Add a new column:**



Categorize the last column and sort the data by that column in DESC:



**Removing Duplicative and Sparse Data**

**df used:**



**Get duplicated rows using .duplicated() function:**  df.duplicated()



It will mark rows that are duplicates.

If you want to see original rows and their duplicates (like in Excel) then add (keep=False):



**Look for duplicates in a specific column using *subset*:**



Or if you want to use multiple columns:



**Delete duplicates** df.drop\_duplicates(inplace=True)

**Delete duplicates based on a column** df.drop\_duplicates(subset=['record\_id'], inplace=True)

Number of unique values in a column df.nunique()



**Identify columns with single-values (e.g. value Shop in the whole column):**



**Drop these columns:** df.drop(delete\_cols, axis=1, inplace=True)

**Read csv file (open csv file) (using a link):**

wine\_df = pd.read\_csv('https://raw.githubusercontent.com/thomasnield/machine-learning-demo-data/master/regression/winequality-red.csv')

wine\_df

**Get number of rows and columns from .shape function:**



**Alternatively. Get number of rows ([0]) and columns ([1]) from the .shape function (where X = df):**



**Count the number of unique values per column:**



**Remove columns with 5% or less unique values:**



**Alternative using scikit-learn, VarianceThreshold and fit\_transform():**



Return from ndarray to get the columns using get\_support:



**Remove columns with duplicates and 3 or less unique values:**



**f-string / f’’ / f”” / printing with f / print(f:**



Example



**Before f’’:**

You can use f’’ like here:

Graphical user interface, text, application

Description automatically generated

**Or you can use %s / %d / %f / %r:**

Graphical user interface, text, application

Description automatically generated

%s is used as a placeholder for the value of name.title(), which is a string. When the print() statement is executed, Python replaces the %s placeholder with the value of name.title(), resulting in a string like "Eric's favorite numbers are:" or "Willie's favorite numbers are:".

The %s is just one of many string formatting operators available in Python, and is used specifically to insert **string** values into a string. Other operators include %d for inserting **integer** values, %f for inserting **floating**-point values, and %r for inserting **any Python** object as a string representation. In Python 3.6 and later versions, a new string formatting method was introduced using f-strings, which offer a more concise and readable way to format strings.

**Handling missing data**

**Looking for missing values**

**Looking for missing values per row** df.isna() *not efficient in my opinion, alternatives:* df.notna(), df.isnull(), df.notnull()



**Check columns for missing values** df.isna().any() *or you can specify axis*  any(axis=0)



**Check rows for missing values** df.isna().any(axis = 1)



**Select (show) columns with NaN (null) values** df.loc[:, df.isna().any()]



**Select (show) rows with NaN (null) values** df.loc[df.isna().any(axis=1), :] *or* df.loc[df.isna().any(axis=1)]



**Look for missing values in specific columns** df[df['TEMPERATURE'].isna() | df['RAIN'].isna()]



**Removing rows with missing values**

Note: many ML and statistical models do not tolerate NA, NaN or other missing null values.

**Remove rows with missing values / drop rows with missing values** df.dropna(axis=0, inplace=True)

**Remove rows with missing values in specific column** df.dropna(axis=0, subset=["RAIN"], inplace=True)

**Remove rows with missing values in specific columns** df.dropna(axis=0, subset=[“TEMPERATURE”, "RAIN"], inplace=True)

**Remove columns with missing values / drop columns with missing values** df.dropna(axis=1, inplace=True)

**Replacing missing values**

**Replace missing values with -1** df.fillna(value=-1, inplace=True)

You can’t specify a column using the subset parameter. To target specific columns you will need to extract them out and then apply fillna().

**NUMPY Replace values with NaN** from numpy import nandf.replace(-1, nan, inplace=True)



**SCIKIT LEARN Replace missing values with mean using SimpleImputer (imputer.fit, imputer.transform)**





There are other options for the *strategy* parameter including 'mean', 'median', 'most\_frequent', and 'constant'.

**SCIKIT LEARN Replace missing values with nearest neighbor using KNNImputer**







**Outliers**

We can use tools like *standard deviation (SD or* *σ [sigma])*  and *interquartile range*.

**Mean**

**Mean** mean = df.mean(axis=0)



**Standard deviation (SD, σ, sigma)**

**Standard deviation** sd = df.std(axis=0)

When calculating standard deviation with Pandas, it will be assumed to be a sample and therefore will calculate with 1 degree of freedom by default as shown in this formula:



To get a sense of how standard deviations play a role in omitting outliers, consider the graphic below. 1 standard deviation away from the mean (average) will capture 68% of the expected data points assuming a normal distribution. 2 standard deviations will capture 95%, and 3 standard deviations will capture 99.7%. With a standard deviation, The lower the standard deviation, the more aggressively outliers will be removed.



For smaller samples, cutting off at two standard deviations will be more common. This means we would declare any data on the tails outside those two standard deviations to be outliers and become candidate for removal.

Let's inspect the outliers outside two standard deviations. Multiply the standard deviation by 2 and subtract/add from the mean respectively to get the lower and upper bounds. Then we can compose a condition to identify the outliers by checking for weights less than or greater than these lower and upper bounds respectively.



**Remove outliers that fall outside the two standard deviations**

df = df[(lower < df[‘column’]) & (df['column'] < upper)]

df

**PANDAS NUMPY Remove outliers that fall outside 2.25 standard deviations:**

Don’t forget the (axis=0)**[0]** for the mean and std.



**Interquartile range outliers**

There is a lot of data that does not follow the nice bell curve shape of the normal distribution. Another way you can approach outliers in these cases is to use the Interquartile Range method, or IQR. This is the difference between the 75th and 25th percentile. When referring to the quarterly percentiles (0, 25, 50, 75, and 100). we refer to them as quartiles. A 50 percent quartile would be the middle-most value (the median), or the average of the two most-centered values.

Using the IQR, you will define a cutoff by a factor 𝑘 below or above the 25th and 75th percentile respectively. A common value for 𝑘 is 1.5, whereas a value of 3.0 would be used for more extreme cutoffs.

**Calculate percentile**

**NUMPY** In Python, we can use the **percentile() function** in NumPy to find a given percentile in a datastet.



**Get IQR:**





As you see above, the k value might be too generous for this dataset if we are looking to remove outliers. Maybe there are not extreme enough outliers in this dataset or this technique is just not warranted. But we can try to experiment lowering that k value to see how low the threshold must be before outliers removed. Below, I find a k value of 1.1 removes an outlier, with an index of 11 and weight of 54.

You can also use this technique on multidimensional data, by specifying an IQR policy for each field you want to target the removal of outliers.

**NUMPY Interquartile range outliers full walkthrough:**







[**LocalOutlierFactor**](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html)

From a machine learning perspective, you can treat outliers as a classification. If they are far away from the rest of the datapoints in a multidimensional space, they can be detected as outliers. However, this becomes less reliable on higher dimensional problems due to curse of dimensionality. By leveraging logic that measures how far neighboring data points are, we can leverage the LocalOutlierFactor.

**SCIKIT LEARN Import LocalOutlierFactor** from sklearn.neighbors import LocalOutlierFactor







**Dates and times**

Check the data type of the date columns using *df.dtypes*. It might be *object* instead of *datetime64*. If a date-column has the right data type assigned then you can extract information from that column. Like day of the week and so on.

**Change data type of the date column (parse data)**  parsed\_col = pd.to\_datetime(df[‘column’])

parsed\_col

**Extract day of week** parsed\_col.dt.dayofweek

**Format dates columns as dates while importing csv:**



**Datetime conversion**

[Dataframe pandas conventions.](https://docs.python.org/3/library/datetime.html#strftime-and-strptime-behavior)

strftime() and strptime() are used to write a datetime to a formatting string, and parse a datetime from a formatted string respectively. The format codes come from the standard C conventions. Here are a few common ones, many of which we will use in this notebook. Refer to the link above to see all format codes.



**ORDER\_DATE\_TM2 column has values like 22-Jan-22 4:08 PM. Correct them to a normal date type:**







**Get all records where day of the week is Tuesday using dt.dayofweek**



**Filter dates**



**Or filter between a specific date and time (option 1)**



And option 2



**Timezones**

**Pytz Library for timezones**

import pytz

pytz.common\_timezones

**Look up a timezone in the pytz library**

tz = pytz.timezone('Europe/Amsterdam')

tz

<DstTzInfo 'Europe/Amsterdam' LMT+0:20:00 STD>

**Check if the column has timezone information using** [**.dt.tz**](https://pandas.pydata.org/docs/reference/api/pandas.Series.dt.tz.html)



**Assign a timezone to a column**



**Convert to a different timezone**



**Convert to UTC**



**Regular expressions (Python *re* library)**

[**re — Regular expression operations**](https://docs.python.org/3/library/re.html)

[**An introduction to regular expressions**](https://www.oreilly.com/content/an-introduction-to-regular-expressions/)

****

Regex quantifiers:

Table

Description automatically generated

[**From:**](https://regex101.com/)

 Graphical user interface, application

Description automatically generated Text

Description automatically generated with medium confidence



[From:](https://regexr.com/)

Graphical user interface, application

Description automatically generated

[Easiest way to remember Regular Expressions (Regex)](https://towardsdatascience.com/easiest-way-to-remember-regular-expressions-regex-178ba518bebd)

**Wrangling text / data cleaning**

These are the common string operations in Pandas we can use. Note that these typically accept a regular expression as a pattern, and we will cover this.

**Function Description *(functions in bold used below)***

count() Counts the number of instances in a pattern, e.g in the following list we have 3 occurrences of number 5:

A screenshot of a computer

Description automatically generated with medium confidence

**contains()** Returns a boolean True/False indicating whether a string contains a pattern

replace() Replaces the found patterns in a string with another specified string.

**fullmatch()** Determines if the entire string matches the pattern

split() Splits a string into separate strings using the pattern as the separator

extract() Finds all occurrences of a pattern and packages them into columns

**findall()** Finds all occurrences of a pattern and packages them into a list

**Fullmatch()**

Graphical user interface, application, Word

Description automatically generated

Same but without coloring:

Graphical user interface, text, application, email

Description automatically generated

**Look for a value in a column.** Look up the value *outlook.com* using the *str.contains()* function in a column. Step 1:

Graphical user interface, text, application

Description automatically generated

**Step 2.** Before showing the values you need to decide on the *NaN* values. If you want them to be treated as *not outlook.com* then add *na=False*. If you don’t add this code you’ll get an error.

Graphical user interface, application

Description automatically generated with medium confidence

**Look for a value in a column alternative.** Using str.fullmatch()

Graphical user interface, text, application, email

Description automatically generated

Show values that don’t match our condition using == False:

Graphical user interface, text, application

Description automatically generated

Only include rows in a dataframe that have a valid phone number and an IP address:

Table

Description automatically generated

Extract email domains from the email column using **str.findall()**



Gather unique domains from one column using str.join() and unique()

Graphical user interface, text, application

Description automatically generated

Show rows that have Eddy as first name:

Graphical user interface, text, application, email

Description automatically generated

Find unique values in a column:

Text

Description automatically generated

**Replace values / replace matched values**



**Splitting columns / splitting text into different columns**

Splitting emails into two columns using str.split

Graphical user interface, text, application, email

Description automatically generated

When you use regular expression features like look-aheads, it opens up more powerful splitting capabilities based on surrounding characters.

Split data in a column into 3 columns (split one column into 3 columns):

Input column:



Code: -, are the symbols used for splitting.

Graphical user interface, text, application

Description automatically generated

**Plot data / Visualize data**

**MATPLOTLIB**

Plot every row in the column weight using .hist()

Chart, histogram

Description automatically generated

**Recursion (countdowns)**

Example 1:

Graphical user interface, text, application

Description automatically generated

Example 2 using if/else/return:

A screenshot of a phone

Description automatically generated with low confidence

**→ Example 3 using a while loop looks like the best option from the 3 examples listed:**

A screenshot of a computer

Description automatically generated with medium confidence

**Fibonacci sequence**

In mathematics, the **Fibonacci sequence is a sequence in which each number is the sum of the two preceding ones**. Numbers that are part of the Fibonacci sequence are known as Fibonacci numbers, commonly denoted Fn . The sequence commonly starts from 0 and 1, although some authors start the sequence from 1 and 1 or sometimes (as did Fibonacci) from 1 and 2. Starting from 0 and 1, the first few values in the sequence are:

0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144.

**Generators (part of iterators / loops / iterations)**

Alternative to iter([1, 2, 3]) is a function with the yield keyword:

A screenshot of a computer

Description automatically generated with low confidence

We can only yield inside a function.

Putting a yield anywhere in a function makes it return generators. Generators are iterators with some more features that we don't need to care about.

Actually we don't even need to use iter() and next() most of the time, but I think it's nice to know how for loops work.

A screenshot of a computer

Description automatically generated with medium confidence

If we have other lines of code between yield then they will be extracted this way:

A screen shot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

[**Intertools**](https://docs.python.org/3/library/itertools.html) **— functions creating iterators for efficient looping.**

For example, itertools.count(1) does the same thing as our count().

A screenshot of a computer

Description automatically generated with low confidence

**Iterable / iterator / for loops**

* An **iterable** is something that we can for loop over.
* An **iterator** is an iterable that remembers its position.
* For loops create an iterator of the iterable and call its \_\_next\_\_ method until it raises a StopIteration.
* Functions that contain yields return generators. Calling next() on a generator runs it to the next yield and gives us the value it yielded.
* [The itertools module](https://docs.python.org/3/library/itertools.html) contains many useful iterator-related things.

**Fill in a list with the help of a for loop / iteration** (add values to a blank list / fill in a blank list [])**:**

A screenshot of a computer program

Description automatically generated with medium confidence

**Comprehensions**

**List comprehensions / dictionary comprehensions / set comprehensions / generator comprehensions (not an official term) / tuple comprehensions / generator expressions**

List comprehensions are a **shorthand way** of creating and working with lists.



**Use list comprehension to shorten this: to this** (an example of a generator expression):

A screenshot of a computer code

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with low confidence

Another example:

A screenshot of a computer program

Description automatically generated with medium confidence

This line can be read as follows: squares = [raise 'number' to the second power, for each 'number' in the range 1-10]

Another example:

A screenshot of a computer code

Description automatically generated with low confidenceA screenshot of a computer

Description automatically generated with medium confidence

Another example:



**Non-numeric comprehensions**

Example 1:

A screenshot of a computer

Description automatically generated with medium confidence

Example 2:

A screenshot of a computer

Description automatically generated with medium confidence

Example 3 with if and else:

A screenshot of a computer code

Description automatically generated with medium confidence

**Generator expression**

**In function calls**

Example 1:

A picture containing text, screenshot, font, line

Description automatically generated

A ‘generator expression’ is this: (ord(symbol) for symbol in symbols).

Example 2:

A screenshot of a computer

Description automatically generated with low confidence

**Example 3:**

Note that in the provided generator expression below values = () and not values = []. We’re not populating an empty list, but

A screenshot of a computer

Description automatically generated with medium confidence

**Cartesian product**

In set theory, the Cartesian product of two sets is the set of all possible ordered pairs where the first element comes from the first set and the second element comes from the second set.

For example, let's say we have two sets A = {1, 2} and B = {x, y}. The Cartesian product of A and B is {(1, x), (1, y), (2, x), (2, y)}. This means that for every element in set A, we can pair it with every element in set B to form a new ordered pair.

Example:

A screenshot of a computer code

Description automatically generated with medium confidence A screenshot of a computer code

Description automatically generated with low confidence A screenshot of a computer

Description automatically generated

Using **format specifiers**, **replacement fields**, and .format():

A screen shot of a computer code

Description automatically generated with low confidence

'<10' and '<5' are called **format specifiers**. They tell the method to format the string to have a field width of 10 characters for color and 5 characters for size.

{:<10} and {:<5} are called **replacement fields**. They indicate where to insert the values of color and size into the formatted string.

**Example 2** with .format():

Here you have two occurrences of {}. They are placeholders. They are filled with information located in the parentheses in the .format() function.

A screenshot of a computer program

Description automatically generated with medium confidence

**Named tuples**

from collections import namedtuple

Since Python 2.6, namedtuple can be used to build classes of objects that are just bundles of attributes with no custom methods, like a database record.

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence

**Pick a random value**

from random import choice

A screenshot of a computer

Description automatically generated with low confidence

**Python Data Model / Dunder Methods**

A screenshot of a computer

Description automatically generated with low confidence

The "Python Data Model" refers to the set of protocols, conventions, and special methods that define how objects in Python behave and interact with each other. It is a conceptual framework that allows Python objects to implement certain operations and behaviors consistently.

The Python Data Model is defined in the Python language documentation and provides a way to customize the behavior of objects in Python by implementing special methods, also known as "magic methods" or "dunder methods" (short for "double underscore" methods). These special methods have predefined names and are used to define how objects respond to various built-in functions, operators, and syntax constructs.

By implementing the appropriate special methods, objects can emulate the behavior of built-in types, support iteration, comparison, arithmetic operations, context management, and more. Examples of these special methods include \_\_init\_\_ for object initialization, \_\_len\_\_ for retrieving the length of an object, \_\_getitem\_\_ for accessing elements using indexing, \_\_str\_\_ for generating a string representation of an object, and many others.

The Python Data Model provides a powerful and flexible way to define custom classes and objects that integrate seamlessly with the rest of the Python language and its features. It allows you to create objects that feel and behave like native Python types, enabling you to write expressive and intuitive code.

The Python Data Model is like a set of rules that define how objects in Python should behave. It tells objects how to respond to certain actions or operations.

Think of it as a guidebook for objects. Just like how people have different behaviors and can respond to various situations, objects in Python can also have different behaviors based on the methods they implement.

For example, if an object implements the \_\_len\_\_ method, it means it can be asked for its length using the len() function. If an object implements the \_\_getitem\_\_ method, it means it can be accessed using indexing or slicing, like a list or a string.

By following the rules of the Python Data Model and implementing the appropriate methods, objects can interact with other parts of the Python language more naturally. It allows you to create custom objects that work seamlessly with built-in functions, operators, and language constructs.

In simpler terms, the Python Data Model helps objects in Python play nicely with the rest of the language by defining how they should behave in certain situations.

**\_\_init\_\_ is called dunder init dunder.**

A picture containing text, screenshot, font, document

Description automatically generated

A screen shot of a computer

Description automatically generated with low confidence

\_\_init\_\_ \_\_str\_\_ \_\_repr\_\_ \_\_len\_\_ \_\_getitem\_\_ \_\_setitem\_\_ \_delitem\_\_ \_iter\_\_ \_\_next\_\_ \_\_eq\_\_ \_\_lt\_\_ \_\_gt\_\_ \_\_add\_\_ \_\_sub\_\_ \_\_mul\_\_ \_\_div\_\_ \_\_truediv\_\_ \_\_call\_\_ \_\_enter\_\_ \_\_exit\_\_ \_\_or\_\_

**Dunder method example class (with \_\_repr\_\_, \_\_init\_\_, \_\_getitem\_\_, \_\_setitem\_\_, \_\_or\_\_, \_\_ior\_\_):**

A screenshot of a computer code

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

dir(dict()):

A screenshot of a computer code

Description automatically generated with medium confidence

\_\_or\_\_:

A screenshot of a computer

Description automatically generated with low confidence

\_\_ior\_\_:

A screenshot of a computer code

Description automatically generated with low confidence

**Overview of magic methods**

**Binary Operators**

|  |  |
| --- | --- |
| **Operator** | **Method** |
| + | `object.\_\_add\_\_(self, other)` |
| - | `object.\_\_sub\_\_(self, other)` |
| \* | `object.\_\_mul\_\_(self, other)` |
| // | `object.\_\_floordiv\_\_(self, other)` |
| / | `object.\_\_truediv\_\_(self, other)` |
| % | `object.\_\_mod\_\_(self, other)` |
| \*\* | `object.\_\_pow\_\_(self, other[, module])` |
| << | `object.\_\_lshift\_\_(self, other)` |
| >> | `object.\_\_rshift\_\_(self, other)` |
| & | `object.\_\_and\_\_(self, other)` |
| ^ | `object.\_\_xor\_\_(self, other)` |
| | | `object.\_\_or\_\_(self, other)` |

A screenshot of a computer

Description automatically generated with low confidence

**Extended assignments**

|  |  |
| --- | --- |
| **Operator** | **Method** |
| += | `object.\_\_iadd\_\_(self, other)` |
| -= | `object.\_\_isub\_\_(self, other)` |
| \*= | `object.\_\_imul\_\_(self, other)` |
| /= | `object.\_\_idiv\_\_(self, other)` |
| //= | `object.\_\_ifloordiv\_\_(self, other)` |
| %= | `object.\_\_imod\_\_(self, other)` |
| \*\*= | `object.\_\_ipow\_\_(self, other[, modulo])` |
| <<= | `object.\_\_ilshift\_\_(self, other)` |
| >>= | `object.\_\_irshift\_\_(self, other)` |
| &= | object.\_\_iand\_\_(self, other) |
| ^= | `object.\_\_ixor\_\_(self, other)` |
| |= | `object.\_\_ior\_\_(self, other)` |

3 examples:

A screenshot of a computer screen

Description automatically generated with low confidence

In each case, the extended assignment operator modifies the value of the left-hand side operand (x in the examples) using the specified operation with the right-hand side operand (y in the examples). The result is then assigned back to the left-hand side operand.

**Unary operators:**

|  |  |
| --- | --- |
| **Operator** | **Method** |
| - | `object.\_\_neg\_\_(self)` |
| + | `object.\_\_pos\_\_(self)` |
| abs() | `object.\_\_abs\_\_(self)` |
| ~ | `object.\_\_invert\_\_(self)` |
| complex() | `object.\_\_complex\_\_(self)` |
| int() | `object.\_\_int\_\_(self)` |
| long() | `object.\_\_long\_\_(self)` |
| float() | `object.\_\_float\_\_(self)` |
| oct() | `object.\_\_oct\_\_(self)` |
| hex() | `object.\_\_hex\_\_(self)` |

Examples:

A screenshot of a computer program

Description automatically generated with medium confidence

A screen shot of a computer program

Description automatically generated with low confidence

**Comparison operators:**

|  |  |
| --- | --- |
| **Operator** | **Method** |
| < | `object.\_\_lt\_\_(self, other)` |
| <= | `object.\_\_le\_\_(self, other)` |
| == | `object.\_\_eq\_\_(self, other)` |
| != | `object.\_\_ne\_\_(self, other)` |
| >= | `object.\_\_ge\_\_(self, other)` |
| > | `object.\_\_gt\_\_(self, other)` |

Examples:

A screenshot of a computer program

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated with low confidenceS

**Example class:**

A screenshot of a computer program

Description automatically generated with low confidence

\_\_repr\_\_ allows to access the object directly using its name like here vector1:

A screenshot of a computer

Description automatically generated with low confidence

\_\_abs\_\_:

A picture containing text, font, screenshot, white

Description automatically generated

\_\_bool\_\_:

A screenshot of a computer

Description automatically generated with low confidence

\_\_add\_\_:

A screenshot of a computer

Description automatically generated with low confidence

\_\_mul\_\_:

A screenshot of a computer

Description automatically generated with low confidence

\_\_call\_\_:

The \_\_call\_\_ method enables Python programmers to write classes where the instances behave like functions. Both functions and the instances of such classes are called **callables**.

**Added on 01.06.2023:**

**Example without Python data model:**

**Class 1:**

A screenshot of a computer program

Description automatically generated with low confidence

Accessing the class:

A screenshot of a computer

Description automatically generated with medium confidence

**Class 2 with data model (note the return instead of print):**

A screenshot of a computer program

Description automatically generated with medium confidence

Accessing the class:

A screenshot of a computer

Description automatically generated with medium confidence

**Functions**

reversed()

A screenshot of a computer

Description automatically generated with medium confidence or A picture containing text, font, white, design

Description automatically generated

**First-class objects**

Functions in Python are **first-class objects**.

Programming language theorists define a **first-class object** as a program entity that can be:

* Created at runtime
* Assigned to a variable or element in a data structure
* Passed as an argument to a function
* Returned as the result of a function

Integers, strings, and dictionaries are other examples of first-class objects in Python — nothing fancy here.

**Difference between return and print**

In Python, return and print serve different purposes.

* return is used within a function to specify the value that the function should evaluate to and return to the caller. When a function encounters a **return** statement, it immediately exits the function and returns the specified value to the caller. The returned value can be assigned to a variable or used in further computations.
* print is used to display output on the console or terminal. It sends the specified values or expressions to the standard output (by default, the console) for display. **print** is typically used for debugging, displaying information, or providing output to the user.

To summarize:

* Use return when you want a function to compute a value and provide it as a result to the caller.
* Use print when you want to display information or output to the console or terminal.

In the context of your \_\_repr\_\_ method, you should use return instead of print to specify the string representation of the object that will be returned when calling repr() on the object.

**Find values that begin or end with a specific value**

Use startswith() or endswith(). Example on the left is preferred (PEP 8 advice) over this (less error prone):



A screenshot of a computer code

Description automatically generated with medium confidence A screenshot of a computer code

Description automatically generated with low confidence

**Compare object types**

Use isinstance to check if your input is equal to a specified datatype. Examples:

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, font

Description automatically generated

Checking something using in:

A close-up of a computer screen

Description automatically generated with low confidence

In this case good is indeed in the text good bye and that is why the result is True.

**replace()**

To replace values with other values use this function (like CTRL + H in Excel):

A screenshot of a computer

Description automatically generated with low confidence

**str.maketrans() и str.translate()**

**Replace a symbol / a list of symbols with a symbol / a list of symbols.**

Используй str.maketrans(), чтобы указать замену буквы/символа. Например внизу *ё* заменяется *е* или во втором примере *a* это *1*, *e* это *2* и так далее. После этого используй str.translate(), чтобы провести замену.

Example 1: Example 2:

A screenshot of a phone

Description automatically generated with low confidence A screenshot of a computer code

Description automatically generated with low confidence

**Exceptions / functions and exceptions**

They are triggered automatically on errors, and they can be triggered and intercepted by your code.

Examples:

1. try/
   1. except Catch and recover from exceptions raised by Python or by you
   2. finally Perform cleanup actions, whether exceptions occur or not
2. raise Trigger an exception manually in your code
3. assert Conditionally trigger an exception in your code
4. with/as Implement context managers

**try/except *(try except, try-except)***

**Example 1:**

It illustrates an option **how to** **cope with errors**. We get a NameError if we run the current code. This can be addressed by using try except.

A screenshot of a computer error

Description automatically generated with low confidence

A picture containing text, font, screenshot, white

Description automatically generated

**Example 2:**

Function that uses a try except:

A screenshot of a computer code

Description automatically generated with low confidence

**Example 3:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer code

Description automatically generated with low confidence

**try/finally**

**Example:**

If a finally clause is included in a try, Python will always run its block of statements “on the way out” of the try statement, whether an exception occurred while the try block was running or not. In its general form, it is:

try:

statements # Run this action first

finally:

statements # Always run this code on the way out

**finally can be seen in the example above (try/except).**

A screenshot of a computer

Description automatically generated with low confidence

**raise**

You can also raise an error inside a function. In the example below an error is raised when an input value is incorrect:

A screenshot of a computer code

Description automatically generated with low confidence

**with/as Context Managers:**

This statement is designed to work with context manager objects, which support a new method-based protocol, similar in spirit to the way that iteration tools work with methods of the iteration protocol.

A screenshot of a computer code

Description automatically generated with low confidence

A picture containing text, screenshot, font, document

Description automatically generated

**PEP 8**

<https://peps.python.org/pep-0008/>

* Use **4 spaces** or a **tab** consisted of 4 spaces
* **Line length** should be 79 for characters and 72 for comments. Alternative is 99 characters. Editors have a setting that shows a vertical line that helps you keep track of the lines.
* Use **blank lines** to break up your code into meaningful blocks. Two can be used in longer programs.
* **Name** variables and program files using only **lowercase** letters, underscores, and numbers.

**Import statements:**

A screenshot of a computer code

Description automatically generated with low confidence

Names of classes can be on the same line:



Imports should always be placed at the top of the file. When you are working on a longer program, you might have an idea that requires an import statement. You might write the import statement in the code block you are working on to see if your idea works. If you end up keeping the import, make sure you move the import statement to the top of the file. This lets anyone who works with your program see what modules are required for the program to work.

Your import statements should be in a predictable order:

* The first imports should be standard Python modules such as sys, os, and math.
* The second set of imports should be "third-party" libraries. These are libraries that are written and maintained by independent programmers, which are not part of the official Python language.

**Errors**

**Ignore errors**

Use %%script false –no-raise-error at the beginning of a code if you don’t want to get an error message / evade an error. Example:

A picture containing text, font, screenshot, white

Description automatically generated

A screenshot of a computer error

Description automatically generated with low confidence

**Most common errors**

Some of the most useful to know and commonly encountered errors include:

* **SyntaxError**: syntax is invalid, such as missing parentheses, mismatched quotes, or incorrect indentation.
* **NameError**: you try to use a variable or name that is not defined or out of scope.
* **TypeError**: operation or function is performed on an object of an inappropriate type. For example, trying to concatenate a string with an integer.
* **IndexError**: you try to access a list, tuple, or other sequence with an index that is out of range.
* **ValueError**: function receives an argument of the correct type but an invalid value. For example, passing a non-integer value to the int() function.
* **KeyError**: you try to access a dictionary with a key that doesn't exist.
* **AttributeError**: you try to access an attribute or method that doesn't exist on an object.
* **FileNotFoundError**: you try to open or access a file that doesn't exist.
* **ImportError**: you try to import a module that cannot be found or accessed.
* **ZeroDivisionError**: you try to divide a number by zero.

**Modules / libraries**

**time module**

**Delay execution time using time.sleep()**

The first line of code will print right away but the second print-statement will execute with a 10 seconds delay:

A screen shot of a computer code

Description automatically generated with low confidence

**itertools module**

Create an itertator that produces the Cartesian product of the input iterables OR simply generate all possible combinations of the elements from the input:

A screenshot of a computer program

Description automatically generated with low confidence

**random module**

**.choices()**

Create a generated list with the values you specified *x* times (k=) long using .choices():



A picture containing text, screenshot, font, line

Description automatically generated

Return 1 random letter from the provided input:

A close-up of a web browser

Description automatically generated with low confidence

Return 1 random number from the provided list:

A picture containing text, font, screenshot, white

Description automatically generated

You can also use the *weight* parameter:

A picture containing text, screenshot, font, line

Description automatically generated

**.sample()**

Using .sample() you can **generate a sample from your data** you already have:

A picture containing text, font, screenshot, line

Description automatically generated

You can also use counts:

(this means that value *1* should occur *10* times and value *5* occurs *1* time: 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5)

A screenshot of a computer

Description automatically generated with low confidence

For numbers you can use range() (5 is k=5):

A picture containing text, font, screenshot, white

Description automatically generated

**.randrange()**

Using .randrange() you can generate a number between the two given numbers:

A close-up of a website

Description automatically generated with low confidence

**collections module**

The Counter class returns a ‘dictionary’ with your given values as keys and the number of these values occurred as values:

A screenshot of a number

Description automatically generated with low confidence

Use .most\_common() to list the most common elements:

A screenshot of a computer code

Description automatically generated with low confidence

You can also add a number in the parenthesis, e.g. .most\_common(3), and that will return top-3 most common elements:

A screen shot of a computer

Description automatically generated with low confidence

**Faker module**

Install and import the module first:

A screenshot of a computer

Description automatically generated with low confidence

If you are looking for data for a specific country you can specify that right away:



Use commands like .name(), .address() to generate fake data:

A screenshot of a computer code

Description automatically generated with low confidence

**pandas module / library**

**DataFrame**

In pandas, a DataFrame is a two-dimensional tabular data structure, similar to a table in a relational database or a spreadsheet. It consists of rows and columns, where each column can have a different data type (e.g., numbers, strings, dates) and each row represents a unique record or observation.

You can think of a DataFrame as a collection of Series objects (one-dimensional labeled arrays) aligned along a common index. Each column in the DataFrame represents a Series, and together they form the table-like structure.

Pandas provides a rich set of functions and methods to manipulate, analyze, and visualize the data in a DataFrame. You can perform operations like filtering rows, selecting specific columns, aggregating data, merging/joining tables, and much more. It's a powerful tool for data manipulation and analysis in Python.

**Create a dataframe using pd.DataFrame({}):**

A screenshot of a computer program

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

Access parts of the dataframe:

A screenshot of a computer

Description automatically generated with medium confidence

**functools module**

**Function reduce**

Apply a function of two arguments cumulatively to the items of a sequence, from left to right, so as to reduce the sequence to a single value. For example, reduce(lambda x, y: x+y, [1, 2, 3, 4, 5]) calculates ((((1+2)+3)+4)+5). If initial is present, it is placed before the items of the sequence in the calculation, and serves as a default when the sequence is empty.

Example 1 using lambda: Example 2:

This example can be easily turned into example 2 by changing the + sign into \*.

A screenshot of a computer

Description automatically generated with low confidence A screenshot of a computer

Description automatically generated with low confidence

**tqdm module**

Use the tqdm() function to **add a progress bar** to visualize the iteration progress.

In the exampe below the function takes an iterable *pages* as an argument and wraps the for loop. Each iteration of the loop updates the progress bar to indicate the progress.

A black and pink rectangular object

Description automatically generated with low confidence

**warnings module**

**warnings.warn**

Use function warnings.warn to **generate a warning message**:

A close-up of a number

Description automatically generated with low confidence

In the example below a ***DepreciationWarning*** message is generated. A *DepreciationWarning* indicates that a Python’s feature or functionality is depreciated and will be removed in the future, and you should probably think of replacing it with something newer. Despite the warning, the code will be executed.

A screenshot of a computer

Description automatically generated with low confidence

**Example 2** with a function *discount* that generates a warning when the discount is higher than 80%:

A screenshot of a computer code

Description automatically generated with low confidence

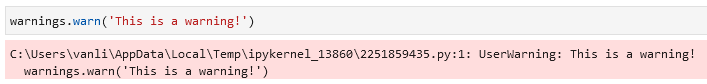
A close-up of a message

Description automatically generated with low confidence

**warnings.simplefilter**

Use function warnings.simplefilter to modify the behavior of how warnings of a specific category are treated. This function has several *actions* like default, ignore, error, etc.

This example shows the functionality of the error *action*. This warning message:



Transforms into an error message:

A screenshot of a computer error message

Description automatically generated with low confidence

By setting the filter to error, the code enforces a stricter behavior regarding ***UserWarning*** and treats them as errors that need to be addressed. This can be useful when you want to ensure that certain warnings are not ignored and are treated as significant issues in your code.

A ***UserWarning*** is a warning class in Python that can be used to issue non-error messages from user code to alert users about potential issues, deviations from best practices, or usage guidance.

Use ignore to ignore all errors and warnings:

A picture containing text, screenshot, font, line

Description automatically generated

Use reset to reset all errors and warnings:

A picture containing text, font, line, screenshot

Description automatically generated

**warnings.catch\_warnings()**

Here are a few scenarios where using warnings.catch\_warnings() can be useful:

1. Ignoring specific warnings: You can use warnings.catch\_warnings() to temporarily ignore specific warnings raised by a specific code section, ensuring that the rest of the code is not affected by those warnings.
2. Suppressing warnings for specific code sections: You might want to suppress all warnings within a particular code section to avoid cluttering the output or to prevent certain warnings from interrupting the flow of the code.
3. Testing warning behavior: During testing, you might want to capture and verify specific warnings raised by a code section. By using warnings.catch\_warnings(), you can capture and inspect the warning messages without letting them propagate and cause test failures.

In general, warnings.catch\_warnings() provides a convenient way to manage warning settings within a specific context. However, it is not necessary to use it in all situations. You can use warnings.simplefilter() to modify the warning behavior globally without the need for a with statement.

**Example that uses a with statement, warnings.catch\_warnings() and warnings.simplefilter():**

You have a function that has a warning:

A white card with red text

Description automatically generated with low confidence

A picture containing text, screenshot, font, line

Description automatically generated

If you want to run the function without the warning:

A screen shot of a computer code

Description automatically generated with low confidence

If you run the function once again you’ll get the same warning:

A picture containing text, screenshot, font, line

Description automatically generated

**\*args and \*\*kwargs**

In the function below my\_arg accepts only 1 value as input, \*args accept any other value that follows, and \*\*kwargs accept all subsequent values in the ‘*dictionary’* style. It is important to follow this structure when writing the function as well as when running the function — values should also be in this order. Otherwise you’ll get an error.

Example 1: Example 2: Example 3: Example 4:

A screenshot of a computer code

Description automatically generated with low confidence A screenshot of a computer code

Description automatically generated with low confidence A screenshot of a computer code

Description automatically generated with low confidence A picture containing text, screenshot, font, line

Description automatically generated

**if, elif, and else in functions**

Encapsulation can be achieved by giving default values (in the example below summer="Play Volleyball", winter="Go snowboarding"), and when running the function you give your own default values:

A screenshot of a calendar

Description automatically generated with medium confidence → A screenshot of a computer program

Description automatically generated with low confidence

**Functions**

**all() function**

all() takes an iterable as its argument and **returns True if all the elements in the iterable are considered truthy** (i.e., non-zero, non-empty, or *True*), and *False* otherwise. It can be used to **check if all the elements in a collection satisfy a certain condition**.

It is also good for memory optimization because it stops when the condition is not met.

Examples:

A screenshot of a computer code

Description automatically generated with low confidence

**enumerate() function**

In addition to helping you out during a loop, you can also use it as a function that adds keys in dictionaries:



A screenshot of a computer

Description automatically generated with low confidence

**Object-oriented programming theory**

Object-oriented programming (OOP) is a programming paradigm that revolves around the concept of objects. In OOP, objects are created from classes, which serve as blueprints or templates for creating objects. Each object is an instance of a class and has its own unique state (data) and behavior (methods).

Here's a breakdown of some key concepts in the object-oriented programming paradigm:

1. **Classes**: Classes are the fundamental building blocks of object-oriented programming. They define the structure and behavior of objects. A class serves as a blueprint or template for creating objects that share common characteristics and functionality.
2. **Objects**: Objects are instances of classes. They represent individual entities that possess unique data (state) and behavior (methods) as defined by their class.
3. **Methods**: Methods are functions defined within a class that operate on the object's data or perform specific actions associated with the object.
4. **Attributes**: Attributes are data variables associated with an object. They represent the object's state or characteristics.

**Matplotlib**

**Sub-libraries (modules) in Matplotlib**

Some commonly used sub-libraries (modules) in Matplotlib:

1. pyplot: provides a MATLAB-like interface for creating plots. It is widely used for creating basic plots, histograms, bar charts, scatter plots, line plots, and more.
2. axes: allows fine-grained control over plot elements and customization of axes, labels, titles, legends, and gridlines. It is commonly used when more advanced customization is needed.
3. figure: provides functionality for creating and managing figures, which are the top-level containers for plots. It allows control over figure size, resolution, background color, and other properties.
4. pylab: combines functionality from both NumPy and pyplot to provide an environment similar to MATLAB. It is convenient for interactive plotting and is often used in interactive environments like Jupyter notebooks.
5. animation: provides tools for creating animated plots and visualizations. It allows you to create dynamic and interactive visualizations with Matplotlib.
6. gridspec: enables the creation of more complex and customized grid layouts for subplots in a figure. It allows you to create multi-row and multi-column subplot configurations.
7. ticker: provides various classes and functions for controlling tick locating and formatting in plots

**Coding styles**

Matplotlib has two coding styles:

* The **pyplot MATLAB-like style** where we reply on pyplot to automatically create and manage the *Figures* and *Axes*.

More explicitly: This style is inspired by MATLAB and provides a convenient way to create and modify plots using functions from the **pyplot** module. The functions in this style directly modify the current active axes or create new axes automatically. It follows a state-based approach where you modify the global state to affect the plot.

* The **object-orient (OO) style** where we explicitly create *Figures* and *Axes* and call methods on them.

More explicitly: This style is based on creating and modifying plot objects directly. It provides a more explicit and flexible way to work with figures and axes. In this style, you explicitly create figure and axes objects, and then call methods on these objects to set various properties of the plot, such as labels, titles, and other customizations. It follows an object-oriented approach where you have more control over individual plot elements and can create complex layouts with multiple subplots.

The **MATLAB-style interface** (**pyplot** interface) is more convenientfor **simple plots or quick visualizations**, while the **object-oriented interface** offers more **flexibility and control** over the plot elements for **complex plots or advanced customization**.

Some commonly used MATLAB-style **pyplot** methods with their equivalent methods in the object-oriented interface

| **MATLAB-style (pyplot)** | **Object-oriented (OO) interface** |
| --- | --- |
| plt.xlabel() | ax.set\_xlabel() |
| plt.ylabel() | ax.set\_ylabel() |
| plt.title() | ax.set\_title() |
| plt.xlim() | ax.set\_xlim() |
| plt.ylim() | ax.set\_ylim() |
| plt.xticks() | ax.set\_xticks() |
| plt.yticks() | ax.set\_yticks() |
| plt.legend() | ax.legend() |
| plt.grid() | ax.grid() or ax.grid(True) |
| plt.plot() | ax.plot() |
| plt.scatter() | ax.scatter() |
| plt.bar() | ax.bar() |
| plt.hist() | ax.hist() |
| plt.imshow() | ax.imshow() |
| plt.colorbar() | fig.colorbar() |

**Importing**

Import pyplot:

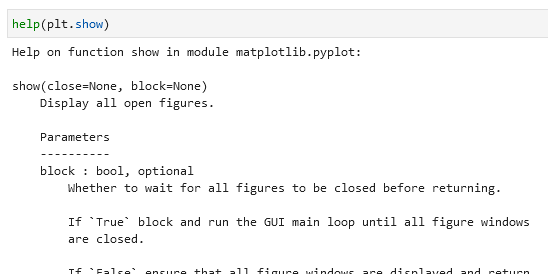


Import ticker:



**Look for help in Matplotlib pyplot**

You can use help() with a function name or you can use the question mark ? after a function name:

 A screenshot of a computer program

Description automatically generated with medium confidence

**See full list of methods for plt., fig., ax.**

Write plt., fig. or ax. and press Tab:

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Filter the result by typing the method name:

A screenshot of a computer program

Description automatically generated with medium confidence

**Older versions**

* Command %matplotlib inline is no longer necessary.

**Anatomy of a figure**

A picture containing text, diagram, number, font

Description automatically generated

**Methods of *Figure* and *Axes* objects**

**Some of the methods of *Figure* object:**

* add\_subplot(): Add a subplot to the figure.
* add\_axes(): Add an axes to the figure.
* savefig(): Save the figure to a file.
* canvas(): Return the FigureCanvas instance.

**Some of the methods of *Axes* object:**

* plot(): Plot lines or markers.
* scatter(): Create a scatter plot.
* bar(): Create a bar plot.
* imshow(): Display an image on the axes.
* set\_xlabel(): Set the x-axis label.
* set\_ylabel(): Set the y-axis label.
* set\_title(): Set the title of the axes.
* legend(): Display a legend on the axes.
* grid(): Display grid lines on the axes.
* set\_xlim(): Set the limits of the x-axis.
* set\_ylim(): Set the limits of the y-axis.

**Return an empty figure**

Use plt.figure() to return an empty figure:

A picture containing text, font, screenshot, white

Description automatically generated

An alternative is to create a chart using e.g. plt.subplots() (see below).

**Create a line-chart using MATLAB style**

Use plt.plot(), plt.xlabel(), plt.ylabel(), plt.title(), and plt.show() to plot a line-chart:

A screen shot of a graph

Description automatically generated with medium confidence A picture containing line, diagram, plot, slope

Description automatically generated

**Create a line-chart using the object-oriented approach**

Use plt.subplots(), ax.plot(), ax.set\_xlabel(), ax.set\_ylabel(), ax.set\_title(), ax.legend(), ax.xaxis.set\_major\_locator(), and plt.show() to plot a line-chart:

* plt.subplots() allows to set the size of the canvas (figure) where the plot will be displayed;
* plt.plot() specifies the x and y coordinates to define the data points for the plot.

**Example with a standard styling:**

A screenshot of a computer program

Description automatically generated with low confidence A picture containing text, line, diagram, plot

Description automatically generated

**Custom styling (including ticker.MaxNLocator):**

A screenshot of a computer program

Description automatically generatedA graph with lines and numbers

Description automatically generated

**Create a scatter plot using the object-oriented approach**

A screen shot of a computer

Description automatically generated with medium confidenceUse plt.scatter to make a scatter plot (for other functions see **line-chart** above):

A picture containing text, screenshot, line, diagram

Description automatically generated

**Create a bar chart using the object-oriented approach**

A screenshot of a graph

Description automatically generated with medium confidenceUse plt.bar to make a bar chart (for other functions see **line-chart** above):

A picture containing text, screenshot, diagram, line

Description automatically generated

**Examples with ticker**

**Show only *integer* values on an axis**

You have this chart:

A screenshot of a computer program

Description automatically generatedA graph with lines and numbers

Description automatically generated

Use ax.xaxis.set\_major\_locator(ticker.MaxNLocator(integer=True) to return only integer values on the x-axis:



 A graph with a line and a point

Description automatically generated

**Show values as an *integer* on an axis**

Using the previous example, if we use ax.xaxis.set\_major(formatter(ticker.FormatStrFormatter(‘%d’)) we will still have 7 ticks on the x-axis, but they will be shown as integers.

 A graph with lines and numbers

Description automatically generated

**Seaborn**

* Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
* [Chart example gallery](https://seaborn.pydata.org/examples/index.html)

**Importing**

Import seaborn:



**Import an** [**example dataset**](https://github.com/mwaskom/seaborn-data)

Seaborn has a number of example datasets. You can load them using the function sns.load\_dataset('dataset\_name'). In the example below we import an example dataset *dots*:

A screenshot of a computer

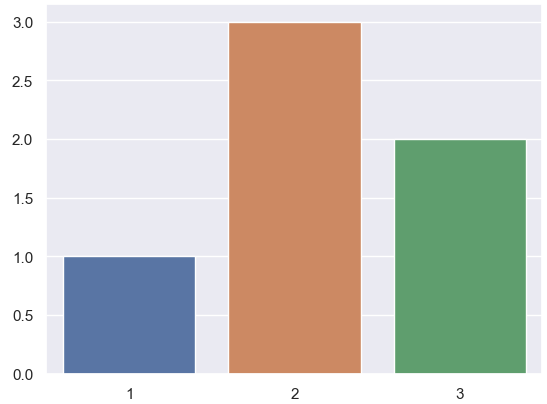
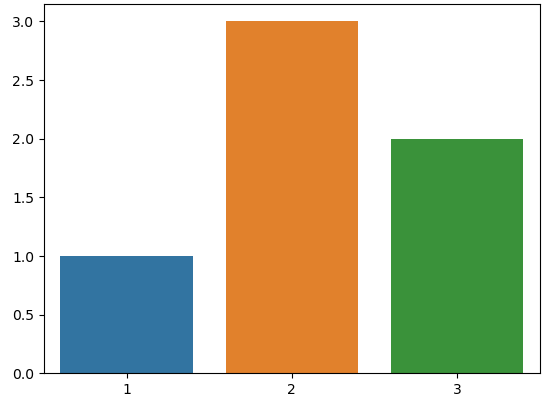
Description automatically generated

**Set deafult Seasborn theme**

Use sns.set\_theme() to set the default theme:

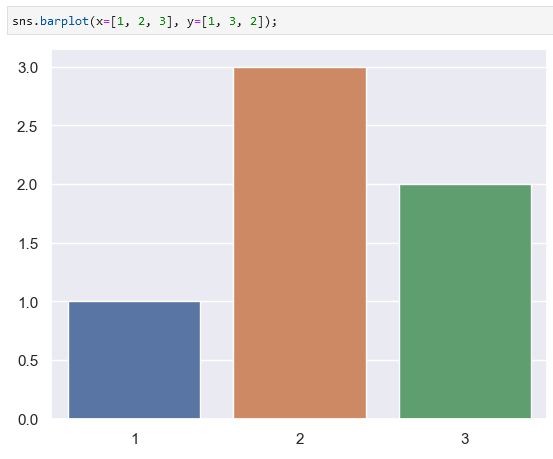


Bar chart with the default theme: Same chart without:

**Plot a bar chart in Seaborn**

Use sns.barplot() to plot a bar chart in Seaborn:



**Plot a histogram in Seaborn**

A screenshot of a computer

Description automatically generated

Use sns.histplot() to plot a histogram in Seaborn. In the example below we use data=df to import the dataframe as data and we specifically select a column for the *x-axis*:

A graph of a graph

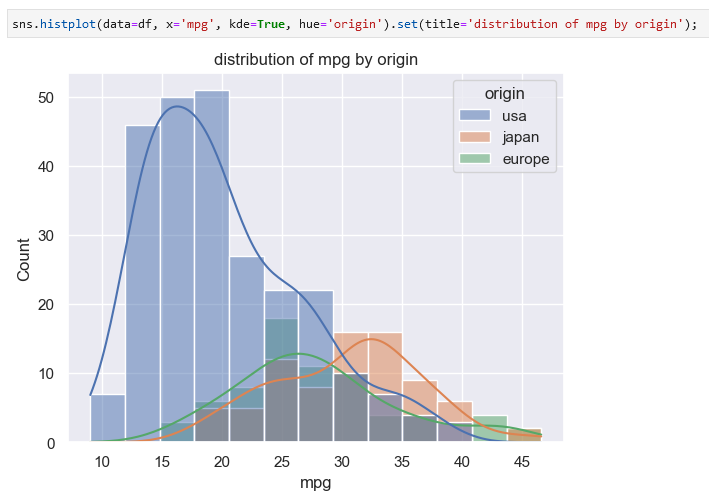
Description automatically generated

Add a smoothed curve that provides an estimate of the underlying distribution of the data by writing kde=True:

A graph of a graph

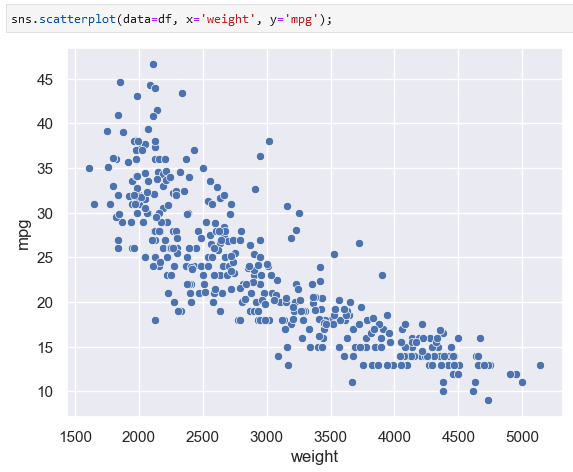
Description automatically generated

If you want to segment the data by a certain factor use hue='column\_name'. In the example below we use column *origin* to segment the histogram by three regions. Add title by using .set(title='title').

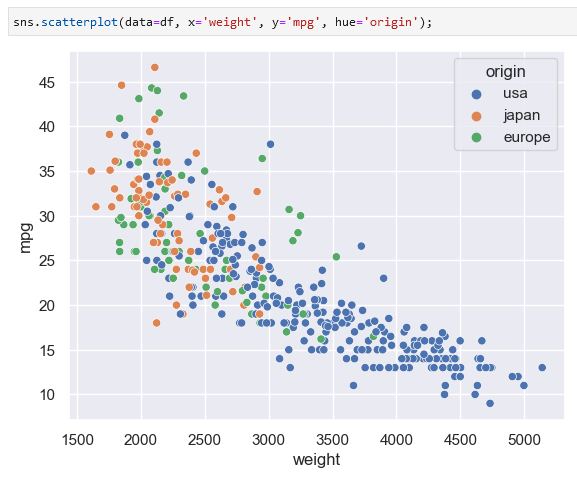


**Plot a scatter plot in Seaborn**

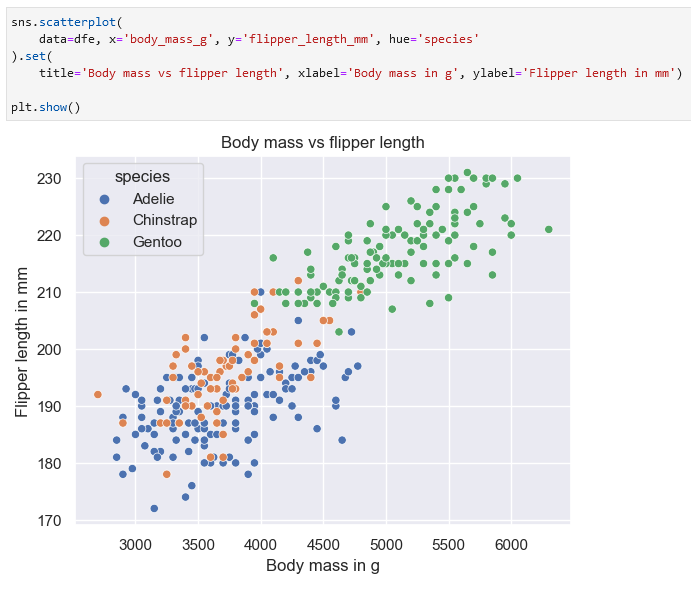
Use sns.scatterplot() to make a scatterplot:



If you want to segment the data by a certain factor use hue='column\_name'. In the example below we use column *origin* to segment the histogram by three regions:



Add title, x-label and y-label name using .set(title='title'), .set(xlabel='name'), .set(ylabel='name'):



**Plot a bar chart based on a column**

This is an alternative to df.column\_name.value\_counts().plot(kind='bar');. Use sns.countplot(data=df, x='column\_name'); to achieve the same result:

A graph with different colored squares

Description automatically generated

**Integration with Matplotlib**

Integrate with Matplotlib by importing it and by using fig, ax = plt.subplots(). You can then use the ax=ax functionality that allows specifying where each plot should go.

A graph with different colored lines

Description automatically generated

In the example below two axes are created ax1 and ax2 so that two graphs can be placed next to each other. Also use nrows and ncols of the plt.subplots function.

A screenshot of a computer

Description automatically generated

**Pandas**

**Importing**

Import pandas:



**Read (open) a csv file in Pandas as a dataframe**

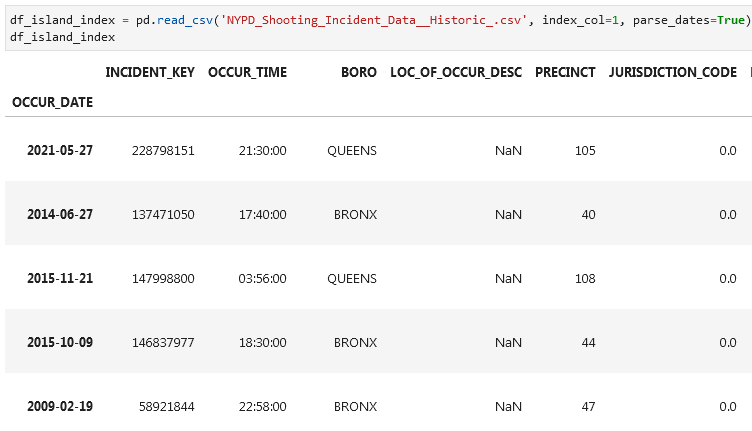
**Standard import**

Use pd.read\_csv() to read (open) a csv file as a dataframe in Pandas. In the example below file name is *mpg* with the *csv* format. The file is then saved as a dataframe df:



**Select a column that will be used as index & parse dates into datetime objects**

Use index\_col=[column number starting 0] to import that column as the index. In case you are using a column that contains dates you can add parse\_dates=True. Pandas will automatically parse dates in the CSV file into datetime objects.



**Create a dataframe (table in Pandas)**

Use pd.DataFrame() to create a dataframe:

A screenshot of a computer

Description automatically generated

**Show first 5 rows in a dataframe**

Use df.head() to show the first 5 rows from a dataframe:

A screenshot of a graph

Description automatically generated

**Get the number of rows in a dataframe**

Use len(df):

A close up of a number

Description automatically generated

**Get basic information on columns in a dataframe like count, mean (average), std, min, max**

Use df.describe():

A screenshot of a graph

Description automatically generated

**Get column labels of a dataframe**

A screenshot of a computer

Description automatically generated

**Select a column of a dataframe**

In the example below we use a column called *weight*.

To select a column type df.column\_name or df['column\_name']

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

**Get the mean of a column (average)**

Select a column using one of the methods and: write df.column\_name.mean() or df['column\_name'].mean().

In the example below we use a column called *weight*:

A screenshot of a computer code

Description automatically generated

**Get the standard deviation of a column**

Select a column using one of the methods and: write df.column\_name.std() or df['column\_name'].std().

In the example below we use a column called *weight*:

A screenshot of a web page

Description automatically generated

**Count values by its occurrence in a column and rank them**

Select a column using one of the methods and: write df.column\_name.value\_counts() or df['column\_name'].value\_counts().

In the example below we use a column called *origin*:

A screenshot of a computer

Description automatically generated

**Perform a GROUP BY…**

A screenshot of a computer

Description automatically generated

**… on the dataframe** Select a column that will be used as labels and write: df.groupby('label\_column\_name').function() (for function check the *… on two columns* example below):

A screenshot of a computer

Description automatically generated

**… on two columns** Use one column as labels and the second column as values that will be summed sum(), averaged mean(), counted count(), minimum min(), and maximum max() values returned etc.

Write: df.groupby('label\_column\_name')['value\_column\_name'].function():

Alternatively:

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

**Plot a standard graph**

Use df.plot() to plot a standard graph of a dataset:

A screen shot of a graph

Description automatically generated

**Plot a histogram based on a column**

**Option 1** Select a column you want to use as input and use df['column\_name'].hist() to plot a histogram. Add grid=True if you want to see a grid like in the example below. In the example below we use a column called *mpg*:

A graph with blue squares

Description automatically generated

**Option 2** Alternative code to plot a histogram is the following df['column\_name'].plot(kind='hist'):



**Create histogram plots for two variables**

**Option 1. On one chart** You can select several columns and plot them as a histogram using this code: df[['column\_name', 'column\_name\_2']].plot(kind='hist'). In the example below we use columns called *PRECINCT* and *JURISDICTION\_CODE*:

A screen shot of a graph

Description automatically generated

**Option 2. Two charts** You can also plot these variables on two separate charts by adding subplots=True:

A screenshot of a graph

Description automatically generated.

**Option 3. Layout** In the example above you had two charts below each other. If you want these charts to be next to each other add layout=(1,n), where n is the number of charts.

A screenshot of a graph

Description automatically generated

**Plot a histogram based on two columns (GROUP BY histogram)**

A screenshot of a computer

Description automatically generated

First select two columns that will have a label & values relationship using df.groupby('label\_column\_name')['value\_column\_name']. Usen then .plot(kind='hist') to plot a graph. If necessary, add legend legend=True and change the transparency of the bars using alpha=n.

A screenshot of a graph

Description automatically generated

**Plot a KDE (Kernel Density Estimate) based on two columns (GROUP BY KDE plot)**

In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function (PDF) of a random variable. This function uses Gaussian kernels and includes automatic bandwidth determination.

Use .plot(kind=kde) to plot this graph.

**One variable:**

A graph of a graph

Description automatically generated

**GROUP BY:**

A graph of a number of data

Description automatically generated

**Plot a bar chart based on a column (look for an easier way in Seaborn)**

To get an overview of a column in the form: labels and the number of occurrence of these labels (see function value\_counts()) use the following code: df.column\_name.value\_counts().plot(kind='bar');.

In the example below figsize=(3, 2) is also used to set the size of the graph.

A graph with blue bars and white text

Description automatically generated or A graph with blue bars and white text

Description automatically generated

A horizontal bar is made using df.column\_name.value\_counts().plot(kind='barh');:

A graph with blue bars and black text

Description automatically generated or A graph with blue bars and white text

Description automatically generated

**Plot a pie chart based on two columns**

To plot a pie chart use the following code: df.column\_name.value\_counts().plot(kind='pie');.

A pie chart with different colors

Description automatically generated A pie chart with different colors and text

Description automatically generated

**Plot a stacked bar chart / an unstacked bar chart based on two columns**

A screenshot of a graph

Description automatically generated

**Unpivot a table**

First you need to unpivot a table using the .unstack() function. Use following code:

df.groupby([column\_1, 'column\_2']).mean()[column\_3].unstack(column\_2)

This table will transform into this:

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

**Creating a stacked bar chart**

Using the previous code as basis add .plot(kind='bar', stacked=True); to plot a stacked bar chart:

A graph of different colored bars

Description automatically generated

**Creating an unstacked bar chart**

Add .plot(kind='bar'); to plot an unstacked bar chart:

A graph of different colored lines

Description automatically generated

**Plot a scatterplot based on two columns / look for a relationship between two columns (variables)**

A screenshot of a graph

Description automatically generated

**Basic scatterplot**

Use the following code to plot a scatterplot: df.plot(x='column\_1', y='column\_2', kind='scatter');. This way you can look for relationships between columns (variables).

A diagram of a graph

Description automatically generated

**Matplotlib Segment the scatterplot by color (Check also in Seaborn [segment the data by a certain factor])**

In case you want to segment your dataframe by color you can use Matplotlib for this.

Use:

import matplotlib.pyplot as plt

fig, ax = plt.subplots()

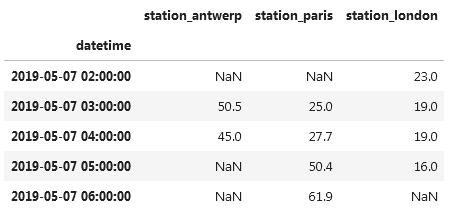
df[df.column\_1=='value'].plot(x='column\_2', y='column\_3', kind='scatter', ax=ax, c='r', label='value');

ax=ax is used to put all of the scatters on one graph. Otherwise every line of code will return on a different graph.

A screen shot of a computer screen

Description automatically generated

**Plot an area chart**



Use plot(kind='area') to plot an area chart:

