Python Introduction

Kevin Sheppard
University of Oxford
www.kevinsheppard.com

September 2019

Contents

Installing

Install Anaconda

1. Download the Anaconda Python/R Distribution 2019.07 (or later).
2. When the download is complete, install into your user account.

Install Visual Studio Code and the Python extension

1. Download VS Code and install
2. Install the Python extension by clicking on Extensions and searching for "Python"
3. Open the `mfe-introduction` folder created in the previous step
4. Create a file called `second.py` and enter

```
#%%

print("Python may be harder to learn than other languages since")
print("there is rarely a single approach to completing a task.")
```

5. Click on Run Cell

Note the `#%%` makes it a magic cell

Install Pycharm Professional

1. Download PyCharm Professional and install using the 30-day trial. You can get a free copy using your academic email address if you want to continue after the first 30 days.
2. Open PyCharm, and create a new project called `mfe-introduction`
3. Open File > Setting and select Python Interpreter. Select the Anaconda interpreter if it is not already selected.
4. Create a new python file called `first.py` and enter

```
print("Python has a steeper curve than MATLAB but more long-run upside")
```

5. Right-click on this file, and select “Run”.

Lesson 1

Getting Started

This lesson covers:

- Opening a terminal window
- Launching Jupyter notebook
- Running IPython in a Terminal
- Running IPython in Jupyter QtConsole
- Executing a standalone Python file in IPython
- Optional
  - Jupyter notebooks in VSCode
  - Jupyter notebooks in PyCharm Professional

1.1 Opening an Anaconda Terminal

An Anaconda terminal allows `python` to be run directly. It also allows other useful programs, for example `pip`, the Python package manager to be used to install packages that are not available through Anaconda.

Windows

Launch Anaconda Prompt from the start menu.

OSX and Linux

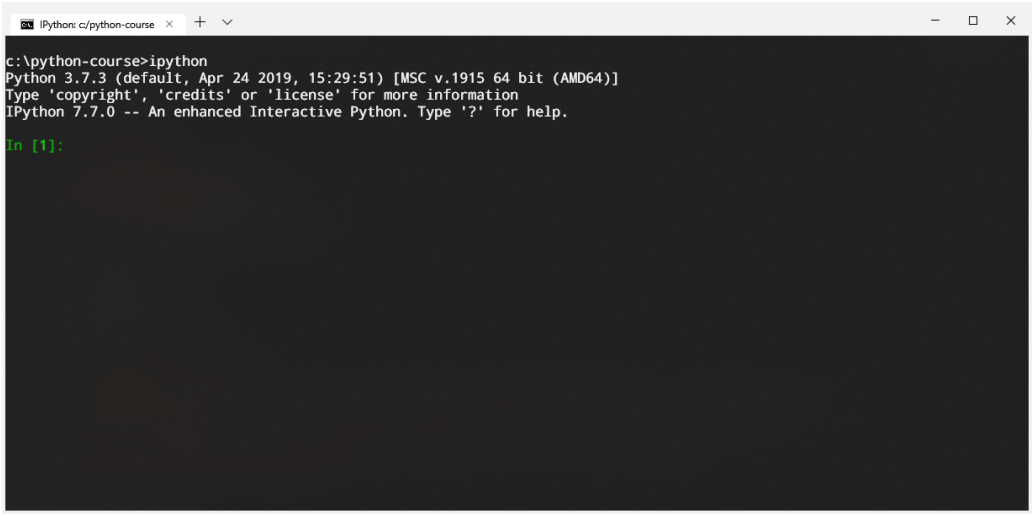
Open the terminal (instructions depend on your distribution). If you allowed conda to initialize, then you should be ready to call Anaconda"s python and supporting functions. If not, you should

```
cd ~/anaconda3/bin
./conda init
```

and then reopen your terminal.

1.2 Running IPython in a Terminal

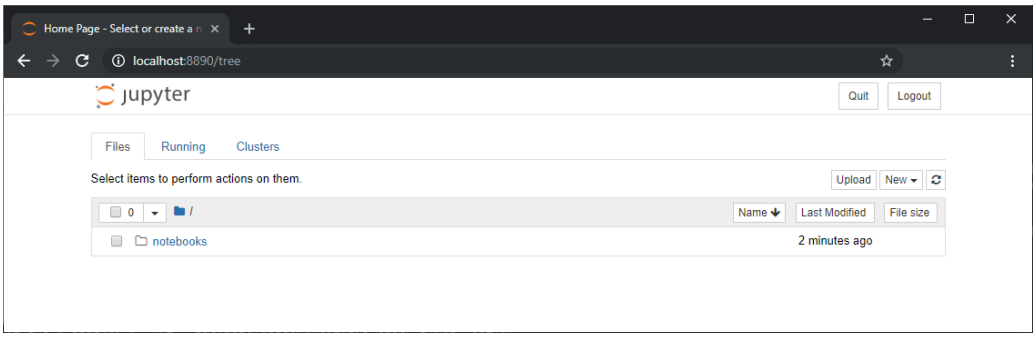
1. Open a terminal.
2. Run IPython by entering `ipython` in the terminal window. You should see a window like the one below with the iconic `In [1]` indicating that you are at the start of a new IPython session.



IPython in Windows Terminal

1.3 Launching Jupyter notebook

1. Launch Jupyter Notebook from the Start Menu or launcher.
2. Change directory to the location where you store your notebooks.



Jupyter Notebook

1.4 Executing a standalone Python file in IPython

1. Open a text editor and enter the following lines. Save the file as lesson-2.py. Note that Python is white-space sensitive, and so these lines should **not** not indented.

```
from math import exp, log

x = exp(1)
y = log(x)

print(f"exp(1)={x}, log(exp(1))={y}")
```

2. Run the code in an IPython session using `%run -i lesson-2.py`. Note: you should create the python file in the same directory as the notebook.

If everything works as expected, you should see

```
exp(1)=2.718281828459045, log(exp(1))=1.0
```

1.5 Jupyter notebooks in VSCode

Visual Studio Code (or VS Code) is a lightweight IDE that supports adding features through extensions. The key extension for working with notebooks is Python extension for Visual Studio Code. With this extension installed, it is possible to use a special file format called Magic Python to write notebook-like files that can be exported to Jupyter notebook files.

1. Install VS Code and the Python extension
2. Open the command palette and enter "Jupyter start" and select the only available item.
3. This is a Python file that supports a cell demarcation using `#%%` for code cells and `#%% [markdown]` for cells that contain markdown code. Note that markdown text **must** be either:

    - Surrounded by triple quotes, e.g. `"""markdown text"""` or `"""markdown text"""`; e.g.,

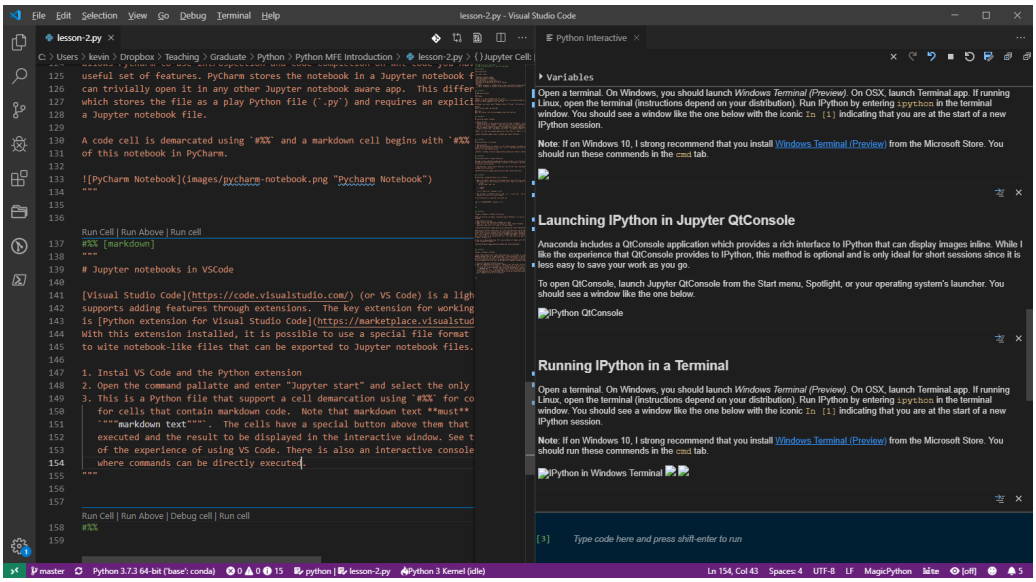
    ```
    """
    # Cell Heading

    Likeness darkness. That give brought creeping. Doesn"t may. Fruit kind
    midst seed. Creature, let under created void god to. Them day was Was
    creature set it from. Fourth. Created don"t man. Man. Light fourth
    light given the he image first multiply after deep she"d great. Morning
    likeness very have give also fowl third land beast from moving thing
    creepeth herb creeping won"t fifth. Us bring was our beast wherein our
    void and green he fruit kind upon a given, saying fruit, moveth face
    forth. His you it. Good beginning hath.
    """
    ```

    - Or commented # (with a single space) at the start of each line,

```
# # Cell Heading
#
# Likeness darkness. That give brought creeping. Doesn"t may. Fruit kind
# midst seed. Creature, let under created void god to. Them day was Was
# creature set it from. Fourth. Created don"t man. Man. Light fourth
# light given the he image first multiply after deep she"d great. Morning
# likeness very have give also fowl third land beast from moving thing
# creepeth herb creeping won"t fifth. Us bring was our beast wherein our
# void and green he fruit kind upon a given, saying fruit, moveth face
# forth. His you it. Good beginning hath.
```

The cells have a special button above them that allows the contents to be executed and the result to be displayed in the interactive window. See the screenshot below for an example of the experience of using VS Code. There is also an interactive console at the bottom left where commands can be directly executed.

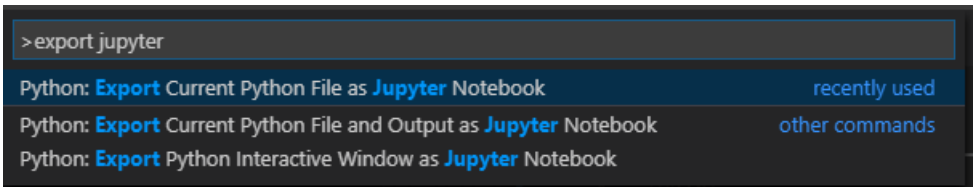


VS Code Notebook

Importing an exiting notebook in VS Code

VS Code only understands Magic Python files as notebook-like documents, and so `.ipynb` files must be converted to use. The process of importing is simple:

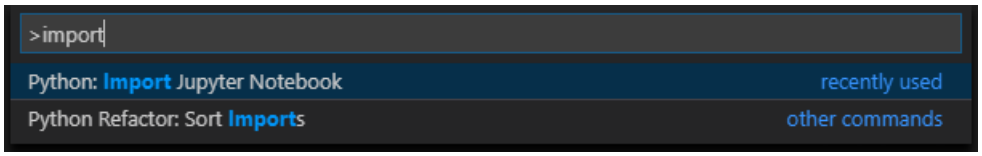
1. Open a Jupyter notebook file
2. Click on Import in the popup that appears.



VS Code Export

Exporting to an Jupyter notebook

To export a Magic Python file, open the command palette and enter "import jupyter". Select the option to import the notebook.



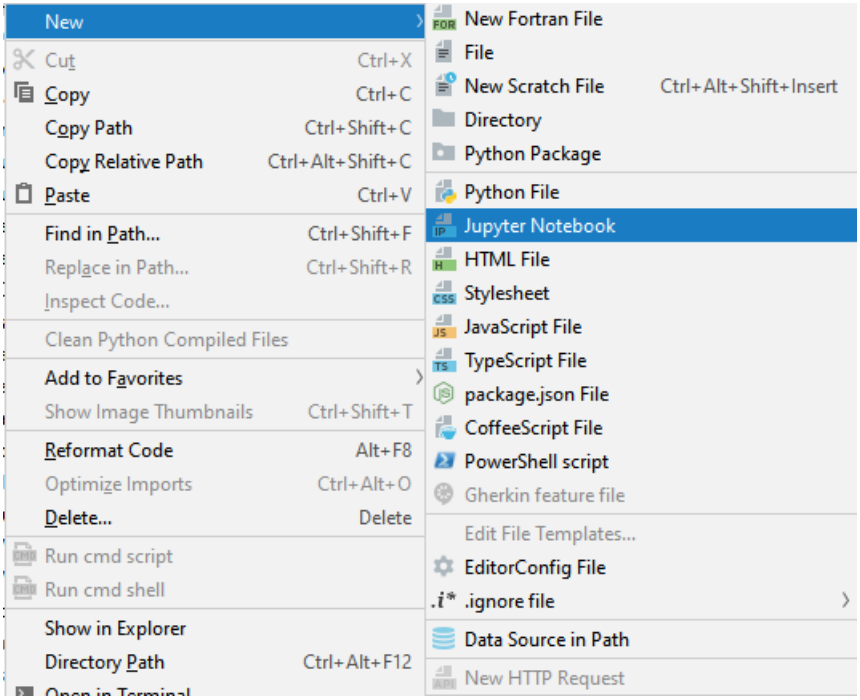
VS Code Import

1.6 Jupyter notebooks in PyCharm Professional

PyCharm Professional is my recommended approach if you are going to use Python throughout the course. It provides the best experience and can be acquired for free using the student program.

PyCharm Professional has deeply integrated Jupyter Notebooks. To create an IPython notebook:

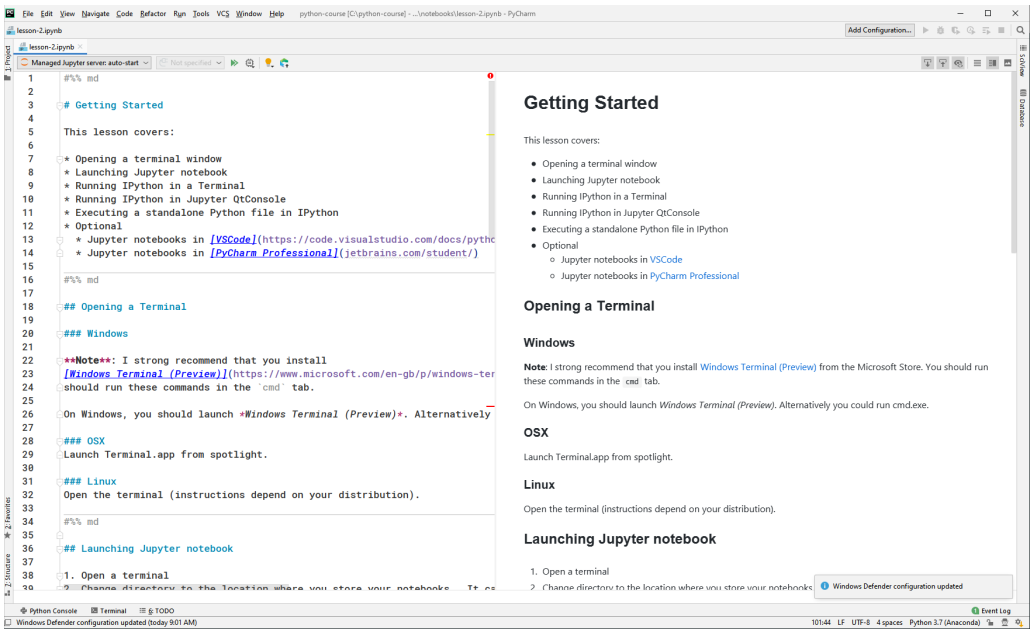
1. Open PyCharm Profession
2. Open the directory where your notebooks are stored
3. Right-click on the root directory and select `New > Jupyter Notebook`. Give your file a meaningful name, and it will open in the main window.



PyCharm New Notebook

PyCharm uses a special syntax where cells look like code and so can be edited like text. This allows PyCharm to use introspection and code completion on the code you have written, a highly useful set of features. PyCharm stores the notebook in a Jupyter notebook file (`.ipynb`), which means that you can trivially open it in any other Jupyter notebook aware app. This differs from Section 1.5 which stores the file as a play Python file (`.py`) and requires an explicit export to a Jupyter notebook file.

A code cell is demarcated using `#%%` and a markdown cell begins with `#%% md`. Below is a screenshot of this notebook in PyCharm.



PyCharm Notebook

Magic Python in PyCharm

PyCharm supports Magic Python cell execution. To use Magic Python, you need to enable *Scientific Mode* in the View menu. You can then use `#%%` to indicate the start and end of cells. Individual Cells can be executed in the console by pressing CTRL+Enter.

1. In PyCharm, right-click on the root directory and select `New > Python File`. Give your file a meaningful name.
2. Enter

```
#%%
print("This is the first cell")

#%%
print("This is not executed when the first cell is run")
```

3. Enable Scientific Mode in the View menu.
4. Run the first cell by placing you mouse in the cell and pressing CTRL+Enter.
5. Run the second cell by clicking on the Play button (arrow) that appears in the gutter of the editor.

Note: Magic Python in PyCharm only supports python code, and so it is not possible to mix Markdown text and Python in the same file.

Lesson 2

Basic Python Types

This lesson covers:

- Inputting scalars and strings
- Lists
- Dictionaries

Problem: Input scalar floating point and integers

1. Create a variable called `scalar_float` containing π to 4 digits.
2. Create a variable called `scalar_int` containing 31415.
3. Print each value using the `print` function.

Problem: Create a string and an f-string

1. Create a variable called `a_string` containing `This is a string`
2. Create a f-string the prints `The value of scalar_float is 3.1415` using the variable created in the previous step
3. Create two string, `first` containing `String concatenation` and the second containing `is like addition`, and join the two using `+` to produce `String concatenation is like addition`.

Problem: Create a list

1. Create a list containing `scalar_float` and `scalar_int`
2. Add `a_string` to the list.
3. Select the the lst two elements of the list

Problem: Create a list of lists

1. Create a list containing the two lists `[1, 2, 3]` and `[4, 5, 6]`
2. Select the element 5 from the nested list

Problem: Create a dictionary

1. Create a dictionary containing the key-value pairs `"float"` and 3.1415, `"int"` and 31415, and `"string"` and `"three-point-one-four-one-five"`.

Problem: Lookup and Change a value

1. Look up the value of `"float"`.
2. Change the value of `"float"` to `22 / 7`.

Problem: Add and remove a key

1. Add the new key “better_float” with the value 3.141592.
2. Remove the key “float” and its value.

Lesson 3

Importing Modules

This lesson covers:

- Module import

Problem: Importing Modules

Python is a general-purpose programming language and is not specialized for numerical or statistical computation. The core modules that enable Python to store and access data efficiently and that provide statistical algorithms are located in modules. The most important are:

- NumPy (`numpy`) - provide the basic array block used throughout numerical Python
- pandas (`pandas`) - provides DataFrames which are used to store data in an easy-to-use format
- SciPy (`scipy`) - Basic statistics and random number generators. The most important submodule is `scipy.stats`
- matplotlib (`matplotlib`) - graphics. The most important submodule is `matplotlib.pyplot`.
- statsmodels (`statsmodels`) - statistical models such as OLS. The most important submodules are `statsmodels.api` and `statsmodels.tsa.api`.

Begin by importing the important modules.

Problem: Canonical Names

Use the `as` keyword to import the modules using their canonical names:

Module	Canonical Name
numpy	np
pandas	pd
scipy	sp
scipy.stats	stats
matplotlib.pyplot	plt
statsmodels.api	sm
statsmodels.tsa.api	tsa

Import the core modules using `import` *module* `as` *canonical*.

Problem: Importing individual functions

1. Import `array`, `sqrt`, `log` and `exp` from NumPy.
2. Import `OLS` from `statsmodels.regression.linear_model`
3. Import the `stats` module from `scipy`

Lesson 4

Series and DataFrames

This lesson covers:

- Constructing pandas Series and DataFrames

Data September 2018 prices (adjusted closing prices) for the S&P 500 EFT (SPY), Apple (AAPL) and Google (GOOG) are listed below:

Date	SPY Price	AAPL Price	GOOG Price
Sept4	289.81	228.36	1197.00
Sept5	289.03	226.87	1186.48
Sept6	288.16	223.10	1171.44
Sept7	287.60	221.30	1164.83
Sept10	288.10	218.33	1164.64
Sept11	289.05	223.85	1177.36
Sept12	289.12	221.07	1162.82
Sept13	290.83	226.41	1175.33
Sept14	290.88	223.84	1172.53
Sept17	289.34	217.88	1156.05
Sept18	290.91	218.24	1161.22
Sept19	291.44	216.64	1158.78

Prices in September 2018

Problem: Input a pandas Series

Create vectors for each of the days in the Section **??** named `sep_xx` where `xx` is the numeric date. For example,

```
import pandas as pd

sep_04 = pd.Series([289.81,228.36,1197.00], index=["SPY","AAPL","GOOG"]);
```

Using the ticker names as the `index` of each series

Problem: Create a Vector of Dates

Use the pandas function `pd.to_datetime` to convert a list of string dates to a pandas `DateTimeIndex`, which can be used to set dates in other arrays.

For example, the first two dates are

```
import pandas as pd

dates_2 = pd.to_datetime(["4-9-2018","5-9-2018"])
print(dates_2)
```

which produces

```
DatetimeIndex(["2018-04-09", "2018-05-09"], dtype="datetime64[ns]", freq=None)
```

Create a vector containing all of the dates in the table.

Problem: Input a Series with Dates

Create vectors for each of the ticker symbols in Section **??** named spy, aapl and goog, respectively. Use the variable `dates` that you created in the previous step as the index.

For example

```
goog = pd.Series([1197.00,1186.48,1171.44,...], index=dates)
```

Set the `name` of each series as the series" ticker.

Problem: Create a DataFrame

Create a DataFrame named `prices` containing Section **??**. Set the column names equal to the ticker and set the index to `dates`.

```
prices = pd.DataFrame([[289.81, 228.36, 1197.00], [289.03, 226.87, 1186.48]],
                      columns = ["SPY", "AAPL", "GOOG"],index=dates_2)
```

Save the price data

This block saves prices to a HDF file for use in later lessons. The function used to save the data is covered in a later lesson.

This function uses some sophisticated ggeatures of Python. Do not worry if it is unclear at this point.

```
# Setup: Save prices, goog and sep_04 into a single file for use in other lessons

# Only run if prices has been defined
if "prices" in globals():
    dates = pd.Series(dates)
    variables = ["sep_04", "sep_05", "sep_06", "sep_07", "sep_10", "sep_11",
                 "sep_12", "sep_13", "sep_14", "sep_17", "sep_18", "sep_19",
```

```
                "spy", "goog", "aapl", "prices", "dates"]
    with pd.HDFStore("data/dataframes.h5", mode="w") as h5:
        for var in variables:
            h5.put(var, globals()[var])
```

Lesson 5

Constructing DataFrames from Series

This lesson introduced method to construct a DataFrame from multiple Series.

This first block loads the variables created in an earlier lesson. A later lesson will cover loading and saving data.

```
# Setup: Load data created in an earlier lesson

import pandas as pd

hdf_file = "data/dataframes.h5"

sep_04 = pd.read_hdf(hdf_file, "sep_04")
sep_05 = pd.read_hdf(hdf_file, "sep_05")
sep_06 = pd.read_hdf(hdf_file, "sep_06")
sep_07 = pd.read_hdf(hdf_file, "sep_07")
sep_10 = pd.read_hdf(hdf_file, "sep_10")
sep_11 = pd.read_hdf(hdf_file, "sep_11")
sep_12 = pd.read_hdf(hdf_file, "sep_12")
sep_13 = pd.read_hdf(hdf_file, "sep_13")
sep_14 = pd.read_hdf(hdf_file, "sep_14")
sep_17 = pd.read_hdf(hdf_file, "sep_17")
sep_18 = pd.read_hdf(hdf_file, "sep_18")
sep_19 = pd.read_hdf(hdf_file, "sep_19")

spy = pd.read_hdf(hdf_file, "spy")
aapl = pd.read_hdf(hdf_file, "aapl")
goog = pd.read_hdf(hdf_file, "goog")

dates = pd.to_datetime(pd.read_hdf(hdf_file, "dates"))

prices = pd.read_hdf(hdf_file, "prices")
```

Problem: Construct a DataFrame from rows

Create a DataFrame named `prices_row` from the row vectors previously entered such that the results are identical to prices. For example, the first two days worth of data are:

```
pricess_row = pd.DataFrame([sep_04, sep_05])
# Set the index after using concat to join
pricess_row.index = dates_2
```

Verify that the DataFrame identical by printing the difference with `prices`

```
print(prices_row - prices)
```

Problem: Construct a DataFrame from columns

Create a DataFrame named `prices_col` from the 3 column vectors entered such that the results are identical to prices.

Note: `.T` transposes a 2-d array since `DataFrame` builds the array by rows.

Verify that the DataFrame identical by printing the difference with `prices`

Problem: Construct a DataFrame from a dictionary

Create a DataFrame named `prices_dict` from the 3 column vectors entered such that the results are identical to prices

Verify that the DataFrame identical by printing the difference with `prices`

Lesson 6

Calling Functions

This lesson covers:

- Calling functions with more than one input and output
- Calling functions when some inputs are not used

Read the data in momentum.csv and creating some variable. This cell uses some magic to automate repeated typing.

```
# Setup: Load the momentum data
import pandas as pd

momentum = pd.read_csv("data/momentum.csv")

print(momentum.head())

mom_01 = momentum["mom_01"]
mom_10 = momentum["mom_10"]
```

This data set contains 2 years of data on the 10 momentum portfolios from 2016–2018. The variables are named mom_XX where XX ranges from 01 (work return over the past 12 months) to 10 (best return over the past 12 months).

Problem: Calling Functions

Functions were used in the previous lesson. Get used to calling functions by computing the mean, std, kurtosis, max, and min of the 10 momentum portfolios.

Use the DataFrame functions `mean`, `std`, `skew` and `kurt` to print the values for `mom_01`.

Problem: Use NumPy and SciPy functions

Use the NumPy functions `mean` and `std` and the SciPy `stats` functions `skew` and `kurtosis` to produce the same output.

Problem: Calling Functions with 2 Outputs

Some useful functions return 2 or more outputs. One example is `np.linalg.slogdet` computes the signed log determinant of a square array. It returns two output, the sign and the log of the absolute determinant.

Use this function to compute the sign and log determinant of the 2 by 2 array:

```
1  2
2  9
```

Problem: Calling Functions with 2 Inputs

Many functions take two or more inputs. Like outputs, the inputs are simply listed in order separated by commas. Use `np.linspace` to produce a series of 11 points evenly spaced between 0 and 1.

Problem: Calling Functions using Keyword Arguments

Many functions have optional arguments. You can see these in a docstring since optional arguments take the form `variable=default`. For example, see the help for `scipy.special.comb`, which has the function signature

```
comb(N, k, exact=False, repetition=False)
```

This tells us that `N` and `k` are required and that the other 2 inputs can be omitted if you are happy with the defaults. However, if we want to change some of the optional inputs, then we can directly use the inputs name in the function call.

Compute the number of distinct combinations of 5 objects from a set of 10.

Compute the total number of combinations allowing for repetition using the `repetition=True` keyword argument.

Compute the number of combinations using the exact representation using only positional arguments for all 3 inputs. Repeat using the keyword argument for `exact`.

Problem: Function Help

Explore the help available for calling functions `?` operator. For example,

```
import numpy as np

np.mean?
```

opens a help window that shows the inputs and output, while

```
import numpy as np

help(np.mean)
```

shows the help in the console.

Lesson 7

Custom Functions

This lesson covers:

- Writing a custom function

Problem: Writing a Custom Function

Custom functions will play an important role later in the course when estimating parameters. Construct a custom function that takes two arguments, mu and sigma2 and computes the likelihood function of a normal random variable.

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Use `def` to start the function and compute the likelihood of:

$$x = 0, \mu = 0, \sigma^2 = 1.$$

The text in the triple quotes is the docstring which is optional.

Exercises

Exercise: Custom Function

Write a function named summary_stats that will take a single input, x, a DataFrame and return a DataFrame with 4 columns and as many rows as there were columns in the original data where the columns contain the mean, standard deviation, skewness and kurtosis of x.

Exercise: Custom Function

Change your previous function to return 4 outputs, each a pandas Series for the mean, standard deviation, skewness, and the kurtosis.

Returning multiple outputs uses the syntax

```
return w, x, y, z
```

Lesson 8

Using DataFrames

This lesson introduces:

- Computing returns (percentage change)
- Basic mathematical operations on DataFrames
- Common DataFrame methods (functions)

This first cell load data for use in this lesson.

```
# Setup: Load prices
import pandas as pd
prices = pd.read_hdf("data/dataframes.h5", "prices")
sep_04 = pd.read_hdf("data/dataframes.h5", "sep_04")
goog = pd.read_hdf("data/dataframes.h5", "goog")
```

Problem: Compute Returns

Compute returns using

```
returns = prices.pct_change()
```

which computes the percentage change.
Additionally, extract returns for each name using

```
spy_returns = returns["SPY"]
```

Problem: Compute Log Returns

```
import numpy as np

log_returns = np.log(prices).diff()
```

first difference of the natural log of the prices. Mathematically this is $r_t = \ln(P_t) - \ln(P_{t-1}) = \ln\left(\frac{P_t}{P_{t-1}}\right) \approx \frac{P_t}{P_{t-1}} - 1$.

Problem: Basic Mathematical Operations

Operation	Symbol	Precedence
Parentheses	()	4
Exponentiation	**	3
Multiplication	*	2
Division	/	2
Floor division	//	2
Modulus	%	2
Matrix multiplication	@	2
Addition	+	1
Subtraction	-	1

Note: Higher precedence operators are evaluated first, and ties are evaluated left to right.

1. Add 1 to all returns
2. Square the returns
3. Multiply the price of Google by 2.
4. Extract the fractional return using floor division and modulus
5. Add the returns on SPY to those of AAPL

Problem: Non-conformable math

Add the prices in `sep_04` to the prices of `goog`. What happens?

Problem: Constructing portfolio returns

Set up a 3-element array of portfolio weights

$$w = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

and compute the return of a portfolio with weight $\frac{1}{3}$ in each security.
Repeat the previous calcuation using multiplication (`*`) and `.sum()`. **Note**: You need to use the `axis` keyword for the sum.

Problem: Mean, Standard Deviation and Correlation

Using the function mean, compute the mean of the three returns series one at a time. For example

```
goog_mean = goog_returns.mean()
```

Next, compute the mean of the matrix of returns using

```
retmean = returns.mean()
```

What is the relationship between these two? Repeat this exercise for the standard deviation (`std()`). Finally, compute the correlation of the matrix of returns (`corr()`).

Problem: Summing all elements

Compute the sum of the columns of returns using `.sum()`. How is this related to the mean computed in the previous step?

Problem: Maximum and Minimum Values

Compute the minimum and maximum values of the columns of returns using the `min()` and `max()` commands.

Problem: Rounding Up, Down and to the Closest Integer

Rounding up is handled by ceil, rounding down is handled by floor and rounding to the closest integer is handled by round. Try all of these commands on 100 times returns. For example,

```
rounded = (100*returns).round()
```

Use `ceil` and `floor` to round up and down, respectively.

Lesson 9

Accessing Elements in DataFrames

This lesson covers:

- Assessing specific elements in Pandas Series and DataFrames

Accessing elements in an array or a DataFrame is a common task. To begin this lesson, clear the workspace set up some vectors and a 5×5 array. These vectors and matrix will make it easy to determine which elements are selected by a command.

Problem: Numeric indexing Series and DataFrame

Repeat the previous questions on `y_s` and `x_df` using `.iloc`.

Problem: Selecting by Name in Series and DataFrames

Using `x_name` and `y_name`:

1. Select the (0,2) and the (2,0) element of `x_name`.
2. Select the 2nd row of `x_name` using `.loc`.
3. Select the 2nd columns of `x_name` using `.loc`.
4. Select the 2nd element of `y_name` using both `[]` and `loc`.
5. Select the 2nd and 4th rows and 1st and 3rd columns of `x_name`.

Problem: Selecting Data by Date

Load the data in momentum.csv.

```
# Setup: Load the momentum data

import pandas as pd

momentum = pd.read_csv("data/momentum.csv", index_col="date", parse_dates=True)
momentum.head()
```

1. Select returns on February 16, 2016.
2. Select return in March 2016.
3. Select returns between May 1, 2016, and June 15, 2016

Lesson 10

Accessing Elements in NumPy Arrays

This lesson covers:

- Accessing specific elements in NumPy arrays

Accessing elements in an array or a DataFrame is a common task. To begin this lesson, clear the workspace set up some vectors and a 5×5 array. These vectors and matrix will make it easy to determine which elements are selected by a command.

Using `arange` and `reshape` to create 3 arrays:

- 5-by-5 array `x` containing the values 0,1,...,24
- 5-element, 1-dimensional array `y` containing 0,1,...,4
- 5-by-1 array `z` containing 0,1,...,4

10.1 Zero-based indexing

Python indexing is 0 based so that the first element has position `0`, the second has position `1` and so on until the last element has position `n-1` in an array that contains `n` elements in total.

Problem: Scalar selection

Select the number 2 in all three, `x`, `y`, and `z`.

Question: Which index is rows and which index is columns?

Problem: Scalar selection of a single row

Select row 2 in `x` and `z` using a single integer value.

Question: What is the dimension of `x` and the second row of `x`

Problem: Slice selection of a single row

Use a slice to select the 2nd row of `x` and the 2nd element of `y` and `z`.

Question: What are the dimension selections?

Problem: List selection of a single row

Use a list to select the 2nd row of x and the 2nd element of y and z.
Question: What are the dimension selections?

Problem: Selecting a single Column

Select the 2nd column of x using a scalar integer, a slice and a list.
Question: What the the dimensions of the selected elemets?

Problem: Selecting Specific Rows or Columns

1. Select the 2nd and 3rd columns of x using a slice.
2. Select the 2nd and 4th rows of x using both a slice and a list.
3. Combine these be combined to select columns 2 and 3 and rows 2 and 4.

Problem: Use `ix_` to select rows and columns using lists

Use `ix_` to select the 2nd and 4th rows and 1st and 3rd columns of x.

Problem: Convert a DataFrame to a NumPy array

Use `.to_numpy` to convert a DataFrame to a NumPy array.

```
# Setup: Create a DataFrame
import pandas as pd
import numpy as np

names = ["a", "b", "c", "d", "e"]
x = np.arange(25).reshape((5,5))
x_df = pd.DataFrame(x, index=names, columns=names)
print(x_df)
```

Problem: Use `np.asarray` to convert to an array

Use `np.asarray` to convert a DataFrame toa NumPy array.

Lesson 11

Numeric Indexing of DataFrames

This lesson covers:

- Accessing specific elements in DataFrames using numeric indices

Accessing elements in a DataFrame is a common task. To begin this lesson, clear the workspace set up some vectors and a 5×5 array. These vectors and matrix will make it easy to determine which elements are selected by a command.

Begin by creating:

- A 5-by-5 DataFrame `x_df` containing `np.arange(25).reshape((5,5))`.
- A 5-element Series `y_s` containing `np.arange(5)`.
- A 5-by-5 DataFrame `x_named` that is `x_df` with columns "c0", "c1", ..., "c4" and rows "r0", "r1", ..., "r4".
- A 5-element Series `y_named` with index "r0", "r1", ..., "r4".

Problem: Picking an Element out of a DataFrame

Using double index notation, select the (0,2) and the (2,0) element of `x_named`.

Problem: Select Elements from Series

Select the 2nd element of `y_named`.

Problem: Selecting Rows as Series

Select the 2nd row of `x_named` using the colon (:) operator.

Problem: Selecting Rows as DataFrames

1. Select the 2nd row of `x_named` using a slice so that the selection remains a DataFrame.
2. Repeat using a list of indices to retain the DataFrame.

Problem: Selecting Entire Columns as Series

Select the 2nd column of `x_named` using the colon (:) operator.

Problem: Selecting Single Columns as DataFrames

Select the 2nd column of `x_named` so that the selection remains a DataFrame.

Problem: Selecting Specific Columns

Select the 2nd and 3rd columns of `x_named` using a slice.

Problem: Select Specific Rows

Select the 2nd and 4th rows of `x_named` using a slice. Repeat the selection using a list of integers.

Problem: Select arbitrary rows and columns

Combine the previous selections to select columns 2 and 3 and rows 2 and 4 of `x_named`.

Note: This is the only important difference with NumPy. Arbitrary row/column selection using `DataFrame.iloc` is simpler but less flexible.

print(x_named.iloc[1:4:2, 1:3]) print(x_named.iloc[[1, 3],[1, 2]]) print(x_named.iloc[[1,3], 1:3])

Problem: Mixed selection

Select the columns c1 and c2 and rows 0, 2 and 4.

Problem: Mixed selection 2

Select the rows r1 and r2 and columns 0, 2 and 4.

Lesson 12

Program Flow

This lesson covers:

- for loops
- Nested loops

Problem: Basic For Loops

Construct a for loop to sum the numbers between 1 and N for any N. A for loop that does nothing can be written:

```
n = 10
for i in range(n):
    pass
```

Problem: Compute a compound return

The compound return on a bond that pays interest annually at rate r is given by $c\,r_t = \prod_{i=1}^{T}(1+r) = (1+r)^T$. Use a for loop compute the total return for £100 invested today for $1, 2, \ldots, 10$ years. Store this variable in a 10 by 1 vector cr.

Problem: Simulate a random walk

(Pseudo) Normal random variables can be simulated using the command `np.random.standard_normal(shape)` where `shape` is a tuple (or a scalar) containing the dimensions of the desired random numbers. Simulate 100 normals in a 100 by 1 vector and name the result `e`. Initialize a vector `p` containing zeros using the function zeros. Add the 1st element of `e` to the first element of `p`. Use a for loop to simulate a process $y_i = y_{i-1} + e_i$. When finished plot the results using

```
%matplotlib inline

import matplotlib.pyplot as plt
plt.rc('figure', figsize=(16,9))
```

```
plt.plot(y)
```

Problem: Nested Loops

Begin by loading momentum data used in an earlier lesson. Compute a 22-day moving-window standard deviation for each of the columns. Store the value at the end of the window.

When finished, plot the annualized percentage standard deviations using `plt.plot(100 * np.sqrt(252) * std_dev)`.

```
# Setup: Load the momentum data

import pandas as pd
momentum = pd.read_csv("data/momentum.csv", index_col="date", parse_dates=True)
momentum = momentum / 100  # Convert to numeric values from percentages
```

Exercises

Exercise

1. Simulate a 1000 by 10 matrix consisting of 10 standard random walks using both nested loops and `np.cumsum`.
2. Plot the results.

Question to think about

If you rerun the code in this Exercise, do the results change? Why?

Lesson 13

Logical Operators

This lesson covers:

- Basic logical operators
- Compound operators

Begin by loading the data in momentum.csv.

```
# Setup: Load the momentum data

import pandas as pd

momentum = pd.read_csv("data/momentum.csv", index_col="date", parse_dates=True)

print(momentum.head())

mom_01 = momentum["mom_01"]
mom_10 = momentum["mom_10"]
mom_05 = momentum["mom_05"]
```

Problem: Basic Logical Statements

For portfolio 1 and portfolio 10, count the number of elements that are < 0, ≥ 0, and exactly equal to 0. Next count the number of times that the returns in portfolio 5 are greater, in absolute value, that 2 times the standard deviation of the returns in that portfolio.

Problem: Compound Statements

Count the number of times that the returns in both portfolio 1 and portfolio 10 are negative. Next count the number of times that the returns in portfolios 1 and 10 are both greater, in absolute value, that 2 times their respective standard deviations.

13.1 Combining Scalar Boolean Values

Use `and`, `or` and `not` to determine if the return on mom_01 on February 14, 2016 is inside or outside an interval of 1 standard deviation of the mean of all of the return of mom_01.

Lesson 14

Conditional Execution

- `if-elif-else` blocks

Problem: Print value if negative

Draw a standard normal value using `np.random.standard_normal` and print the value if it is negative.
Note: Rerun the cell a few time to see different output.

Problem: Print different messages based on value

Draw a standard normal value and print "Positive" if it is positive and "Negative" if not.

Problem:

Draw a standard t random variable with 2 degrees of freedom using `np.random.standard_t(2)` and print "Negative Outlier" if less than -2, "Positive Outlier" if larger than 2, and "Inlier" if between -2 and 2.

Lesson 15

Logic and Loops

This lesson covers:

- Mixing logic and loops

Begin by loading the data in momentum.csv.

```
# Setup: Load the momentum data

import numpy as np
import pandas as pd

momentum = pd.read_csv("data/momentum.csv", index_col="date", parse_dates=True)

print(momentum.head())

mom_01 = momentum["mom_01"]
mom_10 = momentum["mom_10"]
mom_05 = momentum["mom_05"]
```

Problem: Logical Statements and for Loops

Use a for loop along with an `if` statement to simulate an asymmetric random walk of the form

$$y_i = y_{i-1} + e_i + I_{[e_i<0]}e_i$$

where $I_{[e_i<0]}$ is known as an indicator variable that takes the value 1 if the statement in brackets is true. Plot y. e is a standard normal shock. Use `cumsum` to simulate a symmetric one (`z`), and plot the two using the code in the cell below.

Plot the two random walks using the code. We will cover data visualization in a later lesson.

```
%matplotlib inline
import matplotlib.pyplot as plt
plt.plot(y)
```

```
plt.plot(z)
plt.legend(["y", "z"])
```

Problem: Simulate the asymmetricc random walk without an `if-then`

Use boolean multiplication to simulate the same random walk without using an `if-then` statement.

```
# Setup: Plot the data
%matplotlib inline
```

Problem: Selecting Elements using Logical Statements

For portfolio 1 and portfolio 10, select the elements that are < 0, ≥ 0 and exactly equal to 0. Next select the elements where both portfolios are less than 0.

Lesson 16

Using Boolean arrays

This lesson covers:

- `all` and `any`
- `where`

Begin by loading the data in momentum.csv.

```
# Setup: Load the momentum data

import numpy as np
import pandas as pd

momentum = pd.read_csv("data/momentum.csv", index_col="date", parse_dates=True)

print(momentum.head())

mom_01 = momentum["mom_01"]
mom_10 = momentum["mom_10"]
mom_05 = momentum["mom_05"]
```

Problem: Using `where`

Use `where` to select the index of the elements in portfolio 5 that are negative. Next, use the `where` command in its two output form to determine which elements of the portfolio return matrix are less than -2%.

Problem: Combining flow control

For momentum portfolios 1 and 10, compute the length of the runs in the series. In pseudo code,

- Start at i=1 and define run(1) = 1
- For i in 2,...,T, define run(i) = run(i-1) + 1 if sgn (r_i) = sgn (r_{i-1}) else 1.

You will need to use `len` and `zeros`.

1. Compute the length longest run in the series and the index of the location of the longest run. Was it positive or negative?
2. How many distinct runs lasted 5 or more days?

Plot the runs using

```
%matplotlib inline

import matplotlib.pyplot as plt
plt.plot(run)
```

Problem: Use any to find large losses

Use any to determine if any of the 10 portfolios experienced a loss greater than -5%.

Use `all` and negation to do the same check as `any`.

Exercises

Exercise: all and any

Use all to determine the number of days where all of the portfolio returns were negative. Use any to compute the number of days with at least 1 negative return and with no negative returns (Hint: use negation (`~` or `logical_not`)).

Lesson 17

Importing Data

This lesson covers:

- Importing data
- Converting dates

Problem: Reading in data with Dates

Read in the files `GS10.csv` and `GS10.xls` which have both been downloaded from FRED.

Problem: Converting Dates

1. Load the CSV file without converting the dates in `read_csv`.
2. Convert the date column, remove it from the DataFrame, and set it as the index.

Lesson 18

Saving and Exporting Data

This lesson covers:

- Saving and reloading data

This first block loads the data that was used in the previous lesson.

```
# Setup: Load the data to use later
import pandas as pd

gs10_csv = pd.read_csv("data/GS10.csv", index_col="DATE", parse_dates=True)
gs10_excel = pd.read_excel("data/GS10.xls", skiprows=10,
                           index_col="observation_date")
```

Problem: Export to Excel

Export `gs10_csv` to the Excel file `gs10-exported.xlsx`.

Problem: Export to CSV

Export `gs10_excel` to CSV.

Problem: Export to HDF

Export both to a single HDF file (the closest thing to a "native" format in pandas).

Problem: Import from HDF and

Import the data saved as HDF.

Lesson 19

Graphics: Line Plots

This lesson covers:

- Basic plotting
- Subplots
- Histograms
- Scatter Plots

Plotting in notebooks requires using a magic command, which starts with `%`, to initialize the plotting backend.

```
# Setup
%matplotlib inline
```

Begin by loading the data in hf.h5. This data set contains high-frequency price data for IBM and MSFT on a single day stored as two Series. IBM is stored as “IBM” in the HDF file, and MSFT is stored as "MSFT.

Problem: Basic Plotting

1. Plot the `ibm` series which contains the price of IBM.
2. Add a title and label the axes.
3. Add markers and remove the line.

Problem: Subplot

Create a 2 by 1 subplot with the price of IBM in the top subplot and the price of MSFT in the bottom subplot.

Problem: Plot with Dates

Use `matplotlib` to directly plot `ibm` against its `index`. This is a repeat of a previous plot but shows how to use the `plot` command directly.

Lesson 20

Graphics: Other Plots

This lesson covers:

- Histograms
- Scatter Plots

Plotting in notebooks requires using a magic command, which starts with `%`, to initialize the plotting backend.

```
# Setup
%matplotlib inline
```

Begin by loading the data in hf.h5. This data set contains high-frequency price data for IBM and MSFT on a single day stored as two Series. IBM is stored as “IBM” in the HDF file, and MSFT is stored as "MSFT.

Problem: Histogram

Produce a histogram of MSFT 1-minute returns (Hint: you have to produce the 1-minute Microsoft returns first using `resample` and `pct_change`).

Problem: Scatter Plot

Scatter the 5-minute MSFT returns against the 5-minute IBM returns.

Hint: You will need to create both 5-minute return series, merge them, and then plot using the combined DataFrame.

Problem: Saving plots

Save the previous plot to PNG and PDF.

Final Exam

This self-grading notebook serves as a final exam for the introductory course. If you have grasped the contents of the course, you should be able to complete this exam.

It is essential that you answer each cell by assigning the solution to `QUESTION_#` where # is the question number.

We will start with a warm-up question that is already answered.

Question 0

Create a 3-element 1-dimensional array containing the values [1,1,1]

Note: This answer is not assessed.

```
# Setup: The solution is used as a model
import numpy as np

QUESTION_0 = np.ones(3)
```

Question 1

Construct the correlation matrix

$$\begin{bmatrix} 1 & 0.2 & 0.5 \\ 0.2 & 1 & 0.8 \\ 0.5 & 0.8 & 1 \end{bmatrix}$$

as a NumPy array.

Question 2

Construct the correlation matrix

$$\begin{bmatrix} 1 & 0.2 & 0.5 \\ 0.2 & 1 & 0.8 \\ 0.5 & 0.8 & 1 \end{bmatrix}$$

as a DataFrame with columns and index both equal to `['A', 'B', 'C']`.

Question 3

Load the momentum data in the CSV file `momentum.csv`, set the column `date` as the index, and ensure that `date` is a `DateTimeIndex`.

Question 4

Construct a DataFrame using the data loaded in the previous question that contains the returns from momentum portfolio 5 in March and April 2016.

Question 5

What is the standard deviation of the data:

$$1, 3, 1, 2, 9, 4, 5, 6, 10, 4$$

Note Use 1 degree of freedom in the denominator.

Question 6

Compute the correlation matrix of momentum portfolios 1, 4, 6, and 10 as a DataFrame where the index and columns are the portfolio names (e.g., 'mom_01') in the order listed above.

Question 7

Compute the percentage of returns of each of the 10 momentum portfolios that are outside of the interval

$$[\hat{\mu} - \hat{\sigma}, \hat{\mu} + \hat{\sigma}]$$

where $\hat{\mu}$ is the mean and $\hat{\sigma}$ is the standard deviation computed using 1 dof. The returned variable must be a `Series` where the index is the portfolio names ordered from 1 to 10.

Question 8

Import the data the data in the sheet `question 8` in `final-exam.xlsx` into a DataFrame where the index contains the dates and variable name is the column name.

Question 9

Enter the DataFrame in the table below and save it to HDF with the key 'question9'. The answer to this problem must be the full path to the hdf file. The values in index should be the DataFrame's index.

index	data
A	6.0
E	2.7
G	1.6
P	3.1

Note: If you want to get the full path to a file saved in the current directory, you can use

```
import os

file_name = 'my_file_name'
full_path = os.path.join(os.getcwd(), file_name)
```

Question 10

Compute the cumulative return on a portfolio the longs mom_10 and shorts mom_01. The first value should be `1 + mom_10.iloc[0] - mom_01.iloc[0]`. The second cumulative return should be the first return times `1 + mom_10.iloc[1] - mom_01.iloc[1]`, and so on. The solution must be a Series with the name 'momentum_factor' and index equal to the index of the momentum DataFrame.

Note: The data in the momentum return file is in percentages, i.e., a return of 4.2% is recorded as 4.2.

Question 11

Write a function named QUESTION_11 that take 1 numerical input `x` and returns:

- $exp(x)$ is x is less than 0
- $log(1+x)$ if x is greater than or equal to 0

Question 12

Produce a scatter plot of the momentum returns of portfolios 1 (x-axis) and 10 using only data in 2016. Set the x limits and y limits to be tight so that the lower bound is the smallest return plotted and the upper bound is the largest return plotted. Use the 'darkgrid' theme from seaborn. Assign the **figure** handle to QUESTION_12.

Question 13

Compute the excess kurtosis of daily, weekly (using Friday and the end of the week) and monthly returns on the 10 momentum portfolios using the pandas function `kurt`. The solution must be a DataFrame with the portfolio names as the index ordered form 1 to 10 and the sampling frequencies, 'daily', 'weekly', or 'monthly' as the columns (in order). When computing weekly or monthly returns from daily data, use the sum of the daily returns.

Question 14

Simulate a random walk using 100 normal observations from a NumPy `RandomState` initialized with a seed of `19991231`.

Question 15

Defining

```
import numpy as np

cum_momentum = np.cumprod(1 + momentum / 100)
```

compute the ratio of the high-price to the low price in each month. The solution should be a DataFrame where the index is the last date in each month and the columns are the variables names.

Question 16

Simulate 100 observations from the model

$$y_i = 0.2 + 1.2y_{i-1} - 0.2y_{i-2} + \epsilon_i$$

where ϵ_i is a standard normal shock. Set $y_0 = \epsilon_0$ and $y_1 = \epsilon_0 + \epsilon_1$. The solution should be a 1-d NumPy array with 100 elements. Use a RandomState with a seed value of 19991231.

Question 17

What is the ratio of the largest eigenvalue to the smallest eigenvalue of the correlation matrix of the 10 momentum returns?

Note: This is called the condition number of a matrix and is a measure of how closely correlated the series are. You can compute the eigenvalues from the correlation matrix using `np.linalg.eigs`. See the help of this function for more details.

Question 18

Write a function that takes a single input 'x' and return the string "The value of x is" and the value of x. For example, if x is 3.14, then the returned value should be "The value of x is 3.14". The function name must be QUESTION_18.

Question 19

Compute the percentage of days where all 10 returns are positive and subtract the percentage of days where all 10 momentum returns are negative on the same day.

Question 20

Write the function `QUESTION_20` that will take a single input `s`, which is a string and will return a Series that counts the number of times each letter in `s` appears in `s` *without* regard to case. Do not include spaces. Ensure the Series returned as its index sorted.

Hints:

- Have a look at `value_counts` for a pandas `Series`.
- You can iterate across the letters of a string using

```
some_string = 'abcdefg'
for letter in some_string:
    do somethign with letter...
```

- `str.lower` can be used to get the lower case version of a string