Gav McClary

CODING FOR DATA ANALYSIS

An intro to working with Python, Jupyter Notebook and Pandas

OBJECTIVES

- **How to use Jupyter Notebook for data analysis**
- **How to write Python code**
- **Using Pandas library to explore and data**

The aim of this course is to provide you with some fundamental skills to enable you to utilise modern open-source (free!) tools to read, clean, transform and visualise data.

We will start with an introduction to Jupyter Notebook, an open-source web application that allows the creation and sharing of documents that contain live code, equations, visualisations and more.

Next, we will explore the Python programming language and work with some of its basic data structures such as: lists, dictionaries, sets and files.

Following that we will use Pandas library to do some EDA (Exploratory Data Analysis).

I hope you enjoy the course!

Gav

DATA ANALYSIS

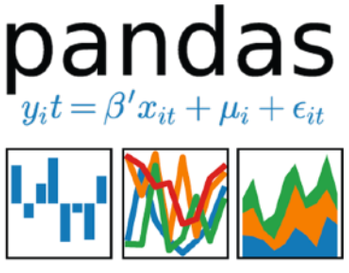
Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making.

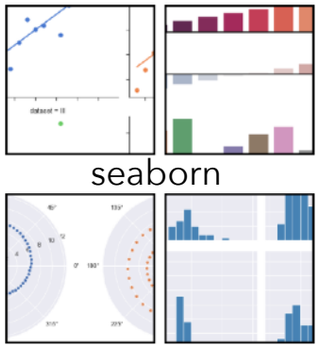
During this course you will learn how to create sample data and also how to load existing data such as **csv** files.

You will then **clean** your data, and transform data in pandas **DataFrames**.

After exploring DataFrames you will **visualise** your data and test some of your new skills!

Why Python?





The reason we use Python for these sessions is because it is a **general-purpose programming language**.

This means it can be used for many other purposes outside of data analysis including:

Web development

GUI (Graphical User Interface) development

Scripting

You may only use it for data analysis for now but the above options become available to you once you understand Python and maybe want to broaden your skill set.

Other domain-specific languages such as **R** are great for statistical work but are not general-purpose languages

Using **Jupyter Notebooks** is a great way to organise and present data analysis and is easy to use and open-source.

The libraries we will use on this course include **Pandas**, **Matplotlib** and **Numpy**. They are also open-source and very popular and powerful packages for data analysis/data science work.

How to use Jupyter Notebook for data analysis

Installation Instructions

To install Jupyter Notebook there are several paths we can take :

Path A: Install Jupyter using Anaconda

Download Anaconda (Python 3.7 version)

Following the instructions on the download page to install

To run search for Anaconda Navigator

Path B: Install Jupyter using pip (experienced Python users only)

```
pip3 install --upgrade pip
```

```
pip3 install jupyter
```

Run Jupyter Notebook with command:

```
jupyter notebook
```

How to use Jupyter Notebook for data analysis

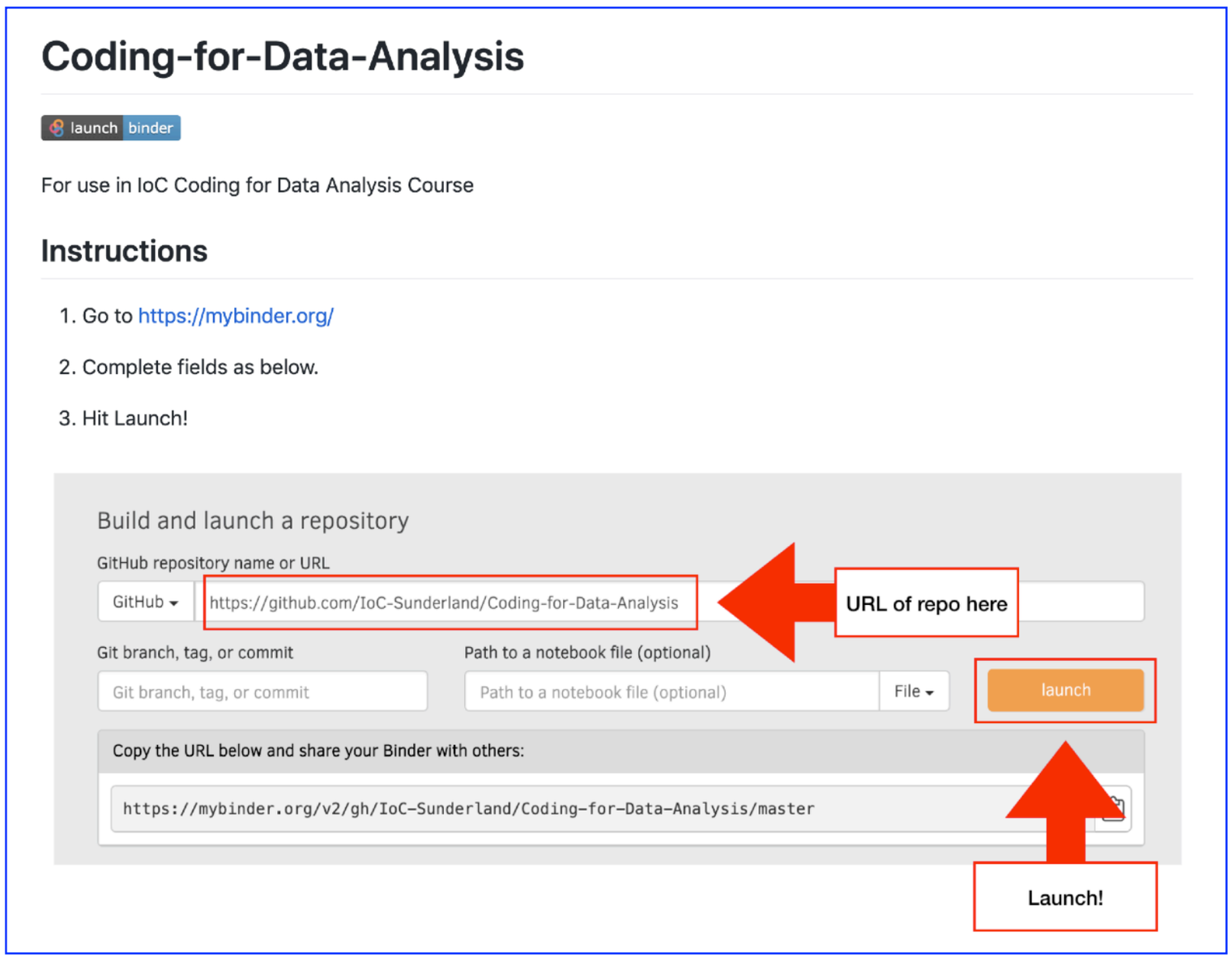
Installation Instructions

Path C: Run Jupyter Notebook using **Binder**

Go to the following url:

https://mybinder.org/v2/gh/IoC-Sunderland/Coding-for-Data-Analysis/master

Or follow instructions as below:



HOW TO USE JUPYTER NOTEBOOK FOR DATA ANALYSIS

What is Jupyter Notebook?

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more (https://jupyter.org).

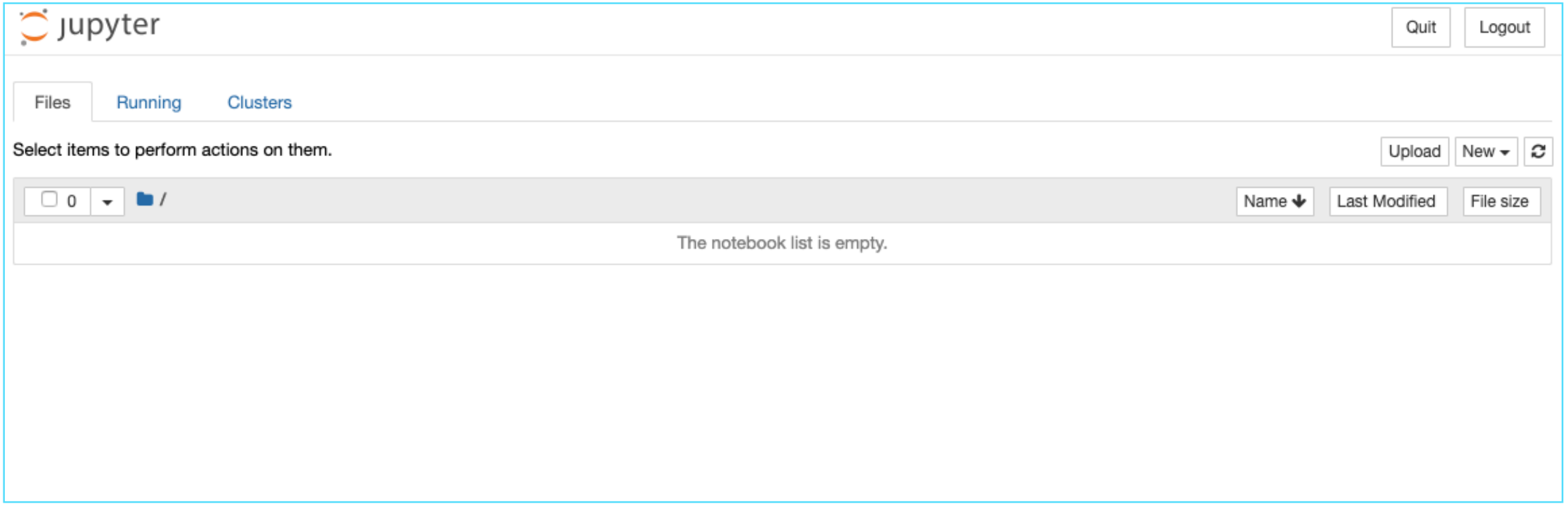
Throughout the course we will be using Jupyter Notebook to execute (run) all the Python code and pandas code so as to keep all our experiments in one place.

We can use Jupyter's **markdown** cells to add comments and descriptions and we can place **visualisations** such as **Box Plots** and **Histograms** inline with our code experiments.

How to use Jupyter Notebook for data analysis

Creating Notebooks

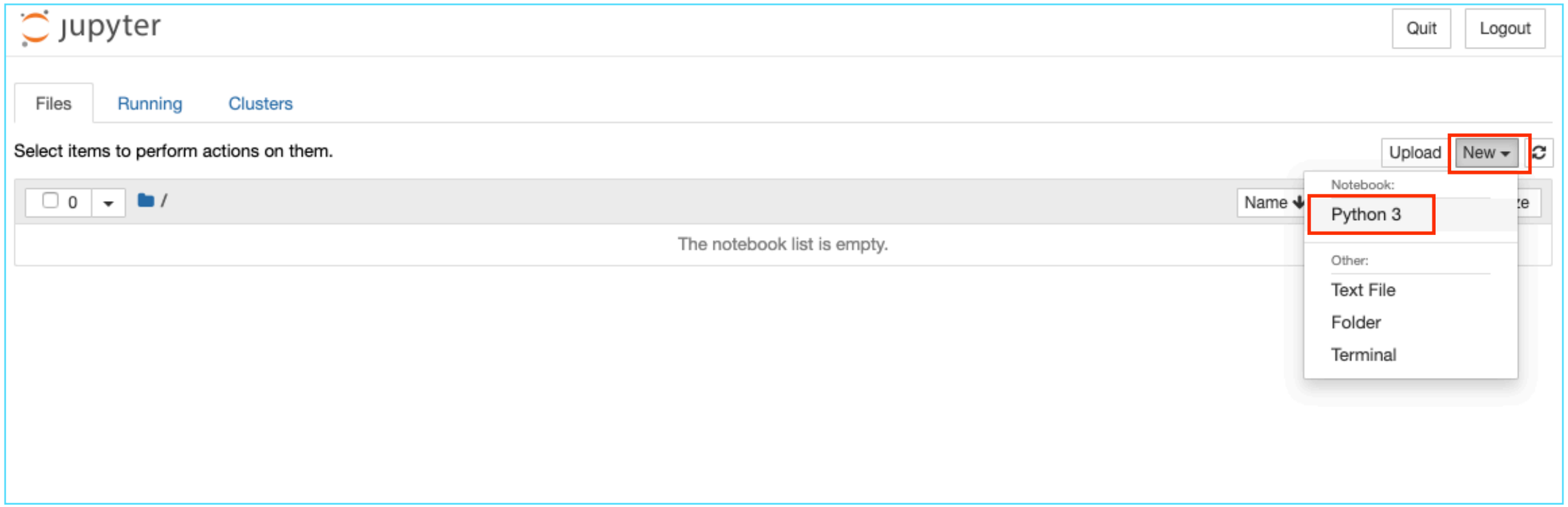
After following the installation steps and running command ***jupyter notebook***
you should be presented with a web page that looks **similar** to this one:



How to use Jupyter Notebook for data analysis

Creating Notebooks

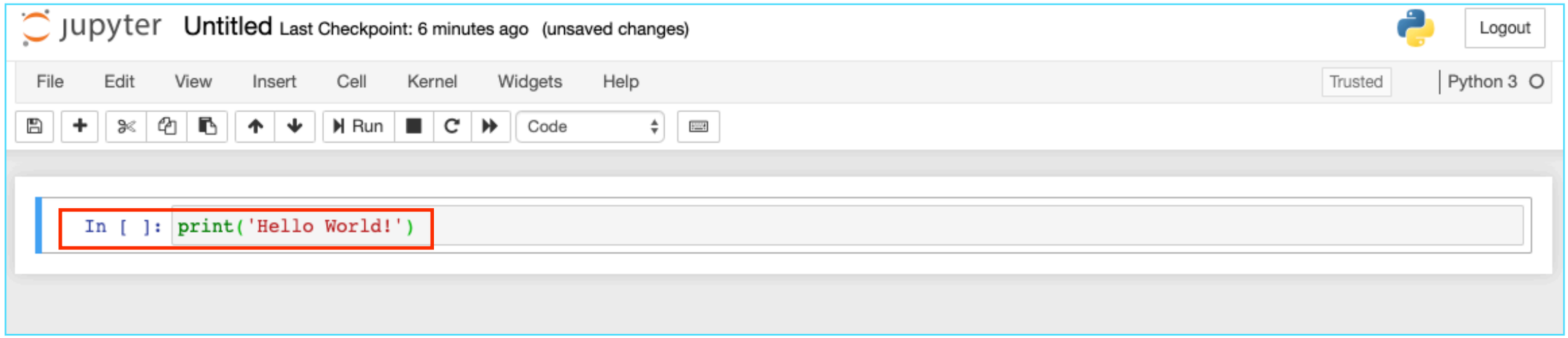
To open a **new notebook** click on **New** then **Python 3**:

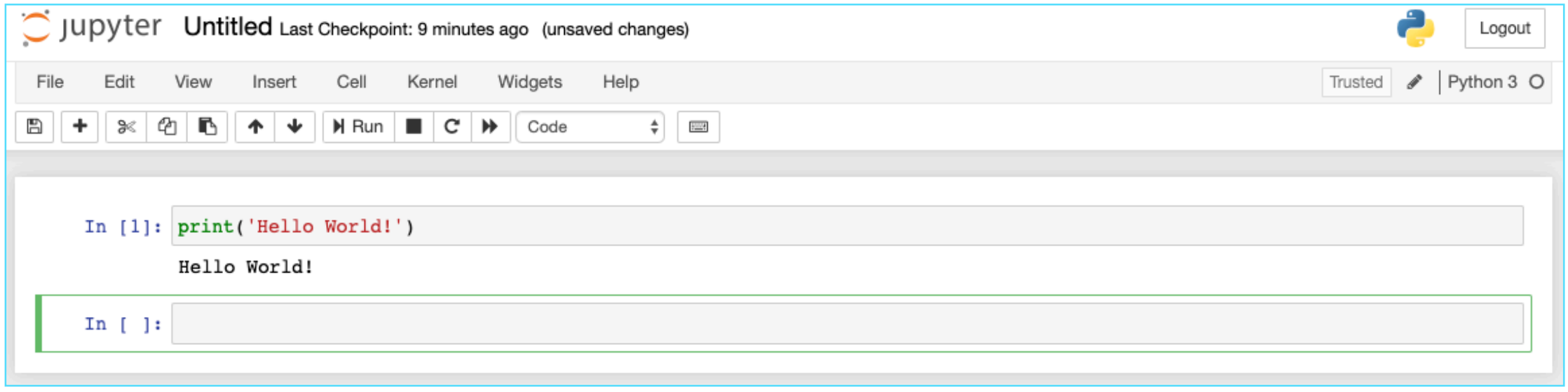


How to use Jupyter Notebook for data analysis

Running Cells

To run a cell, first type some code (see example below) then press the key combination **shift** and **enter** to run that code

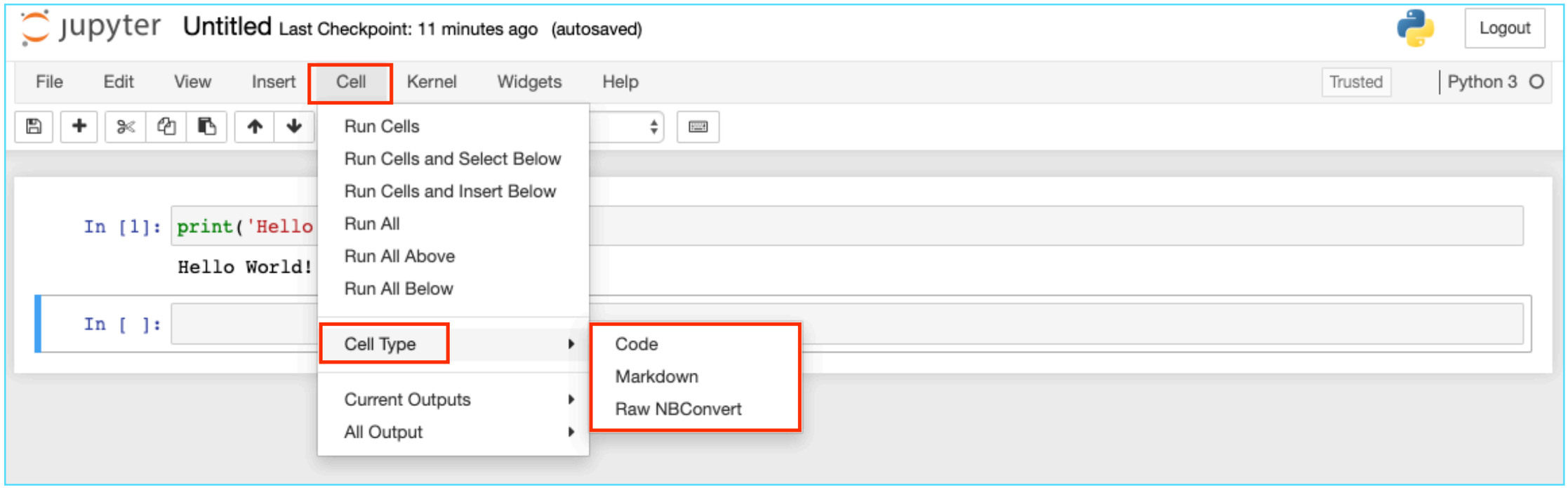




How to use Jupyter Notebook for data analysis

Cell Types

Cells in a notebook can be of a different type. During the course we will be using both the **code** cell type and the **markdown** cell type.



More about **cell types**:

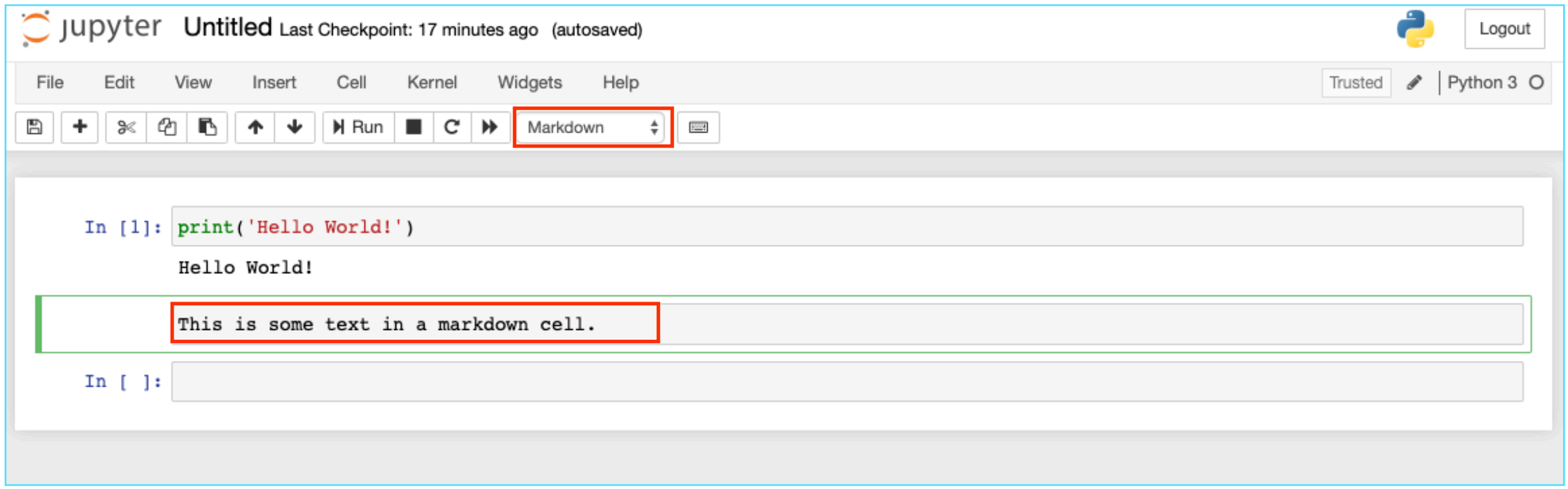
The **code** cell type is used when you want to execute/run some Python or pandas code in the notebook

The **markdown** cell type is used when you want to enter some plain text into the notebook (see example on the next page)

How to use Jupyter Notebook for data analysis

Markdown cells

Use cells of type **markdown** to enter explanatory text into your notebook:



More about **markdown**:

Markdown is a markup language (think a light version of HTML) often used to write web documents.

For some guidance on markdown syntax go here:

https://www.markdownguide.org/basic-syntax

HOW TO WRITE PYTHON CODE

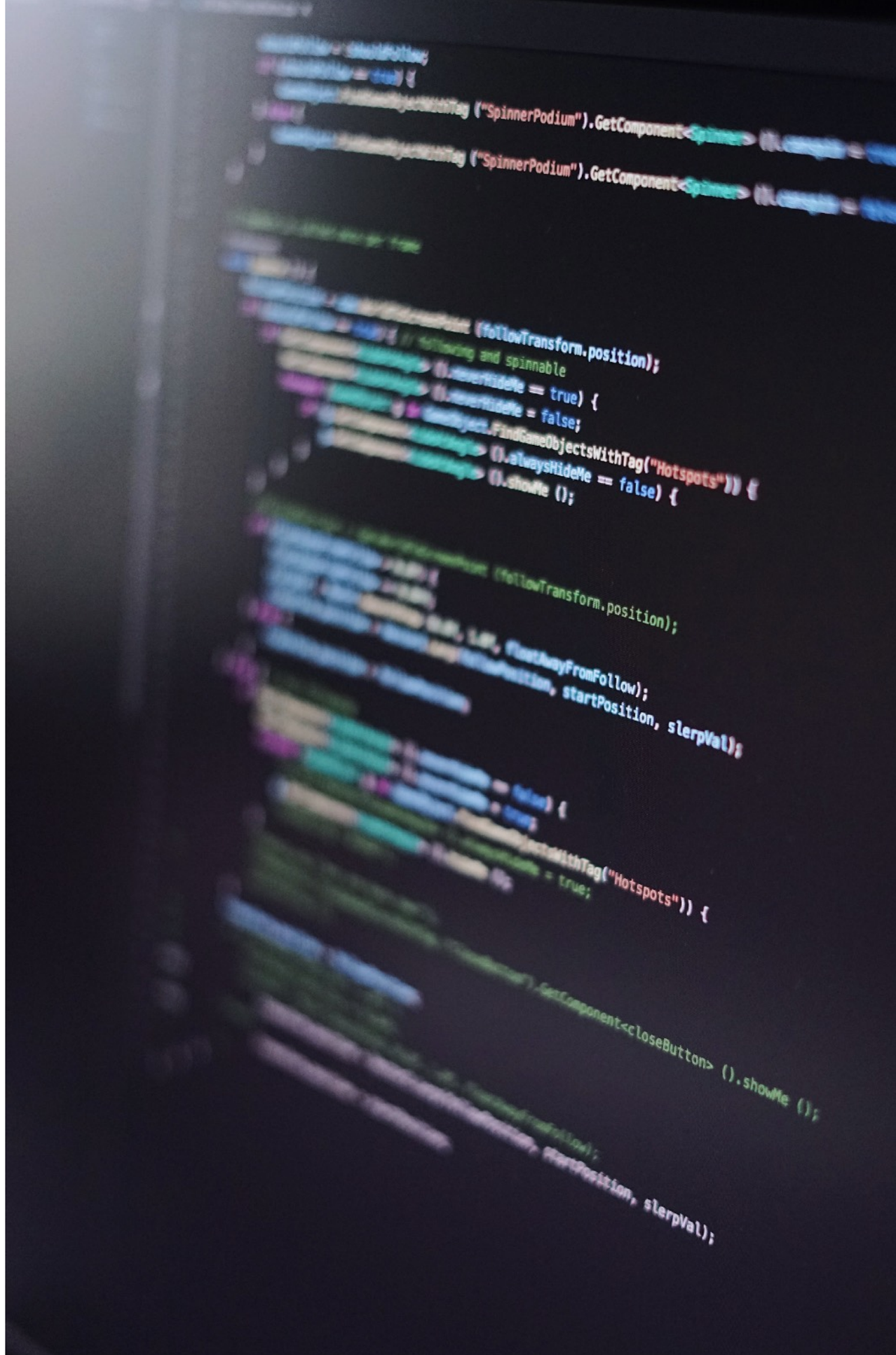
What is Python?

Python is a high-level programming language.

As far as programming languages go it is very human readable and therefore easier to learn than other languages.

Python is commonly used in the **Data Analysis/Data Science** fields and as such has access to many external libraries that support these activities including **NumPy, Matplotlib** and **Pandas.**

Here, we will explore some of Python's basic syntax, data structures and how to work with files.

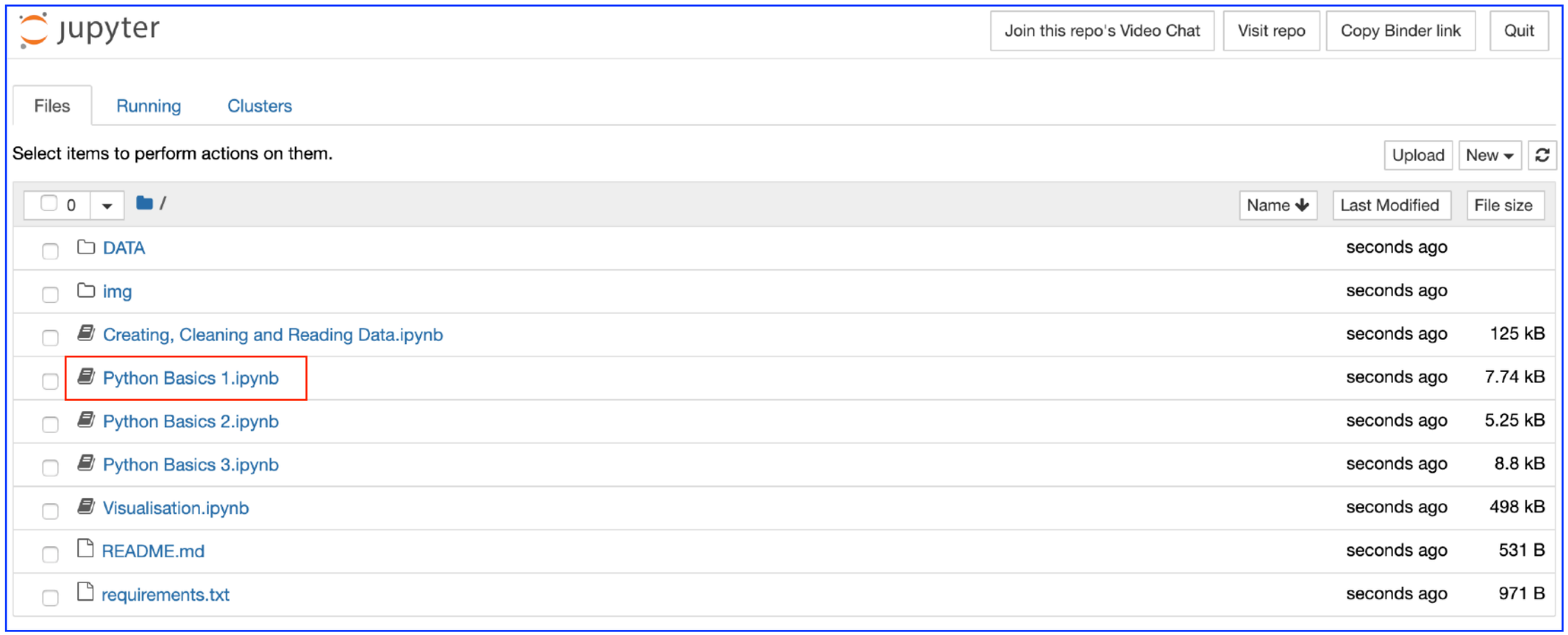


How to write Python code

Using Jupyter Notebook to run Python commands

Python Basics 1

Open the file **Python Basics 1.ipynb** by browsing for the file in Jupyter Notebook and clicking on the link

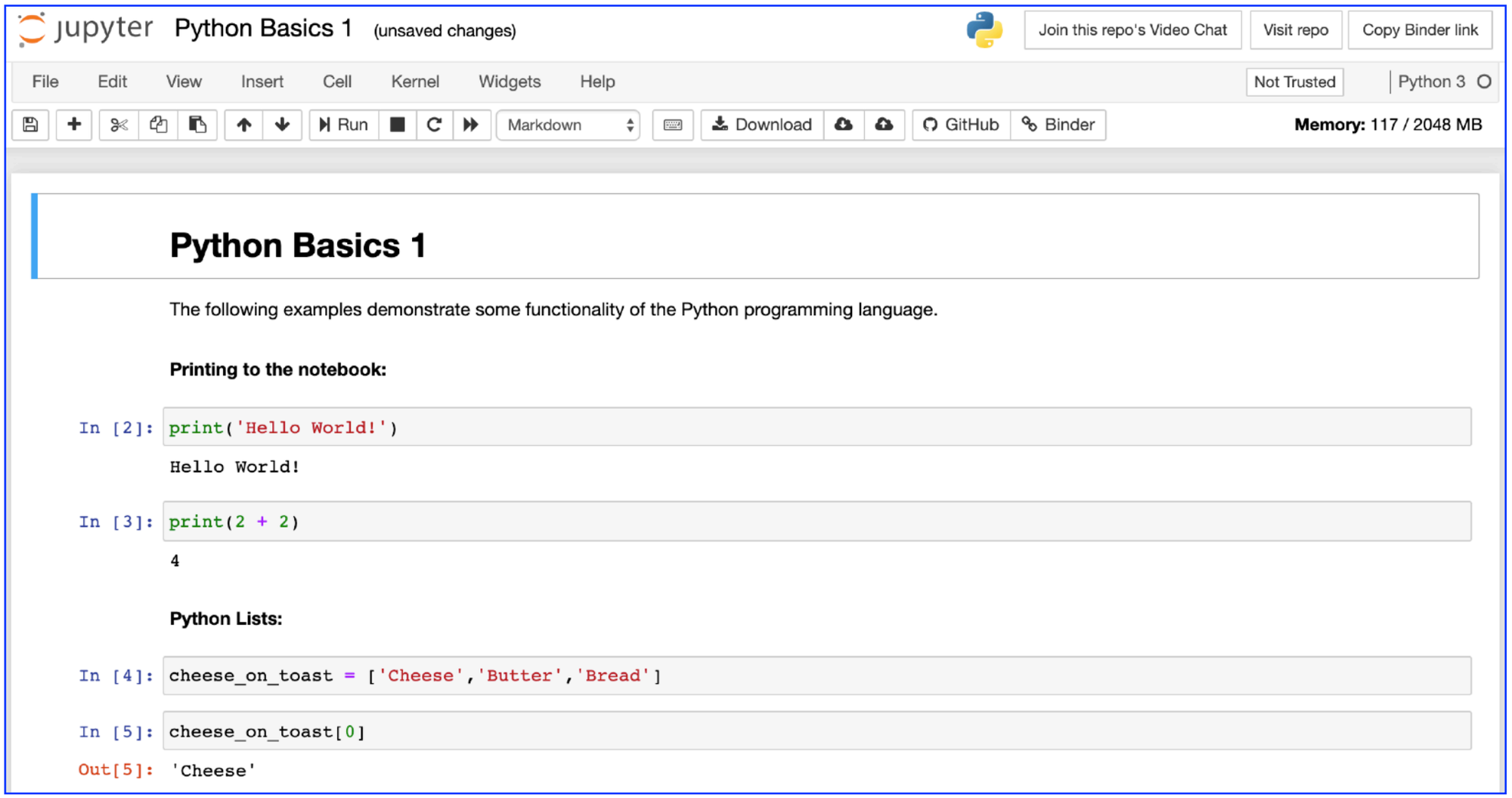


How to write Python code

Using Jupyter Notebook to run Python commands

Python Basics 1

It should display on screen like below:

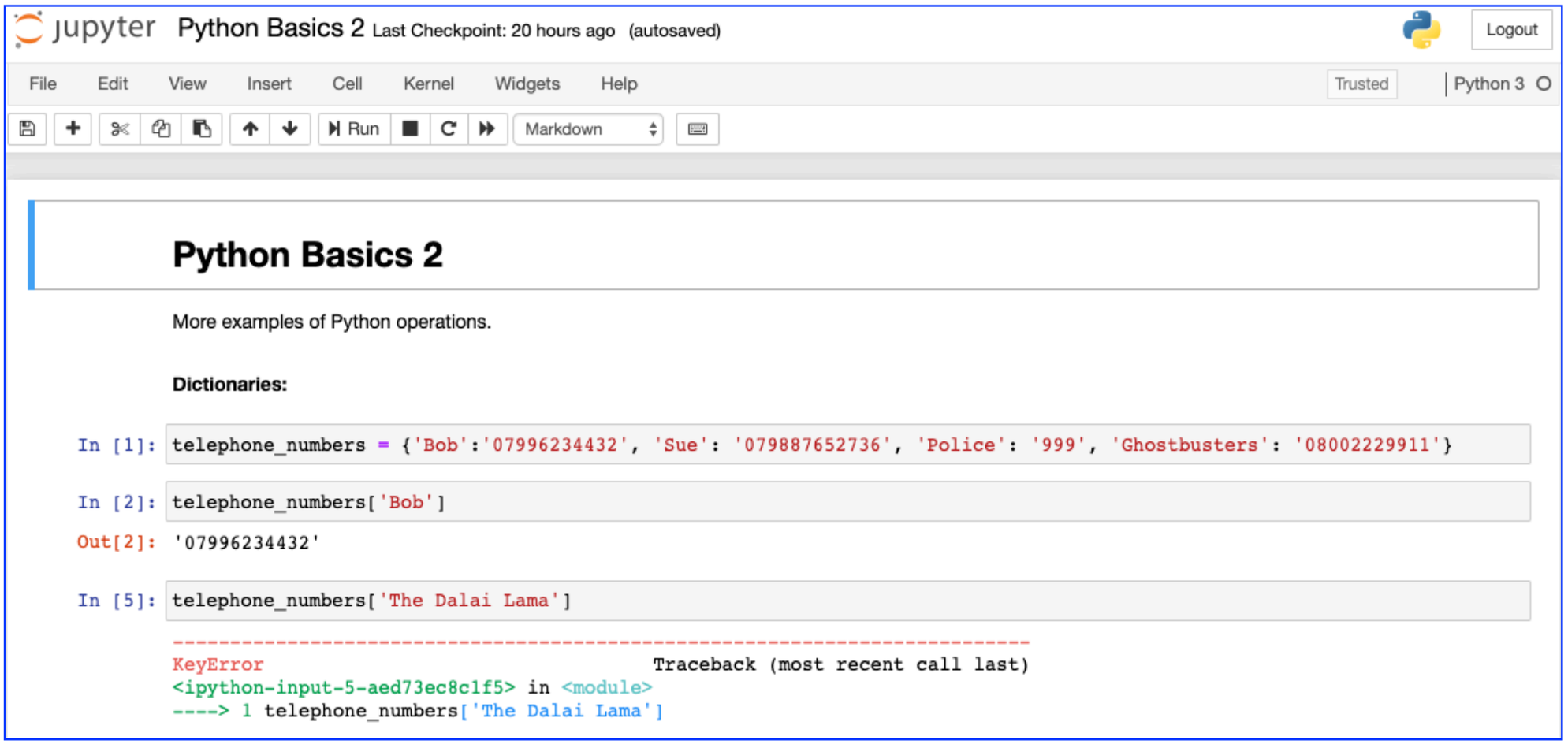


How to write Python code

Using Jupyter Notebook to run Python commands

Python Basics 2

Open the file **Python Basics 2.ipynb** by browsing for file in Jupyter Notebook

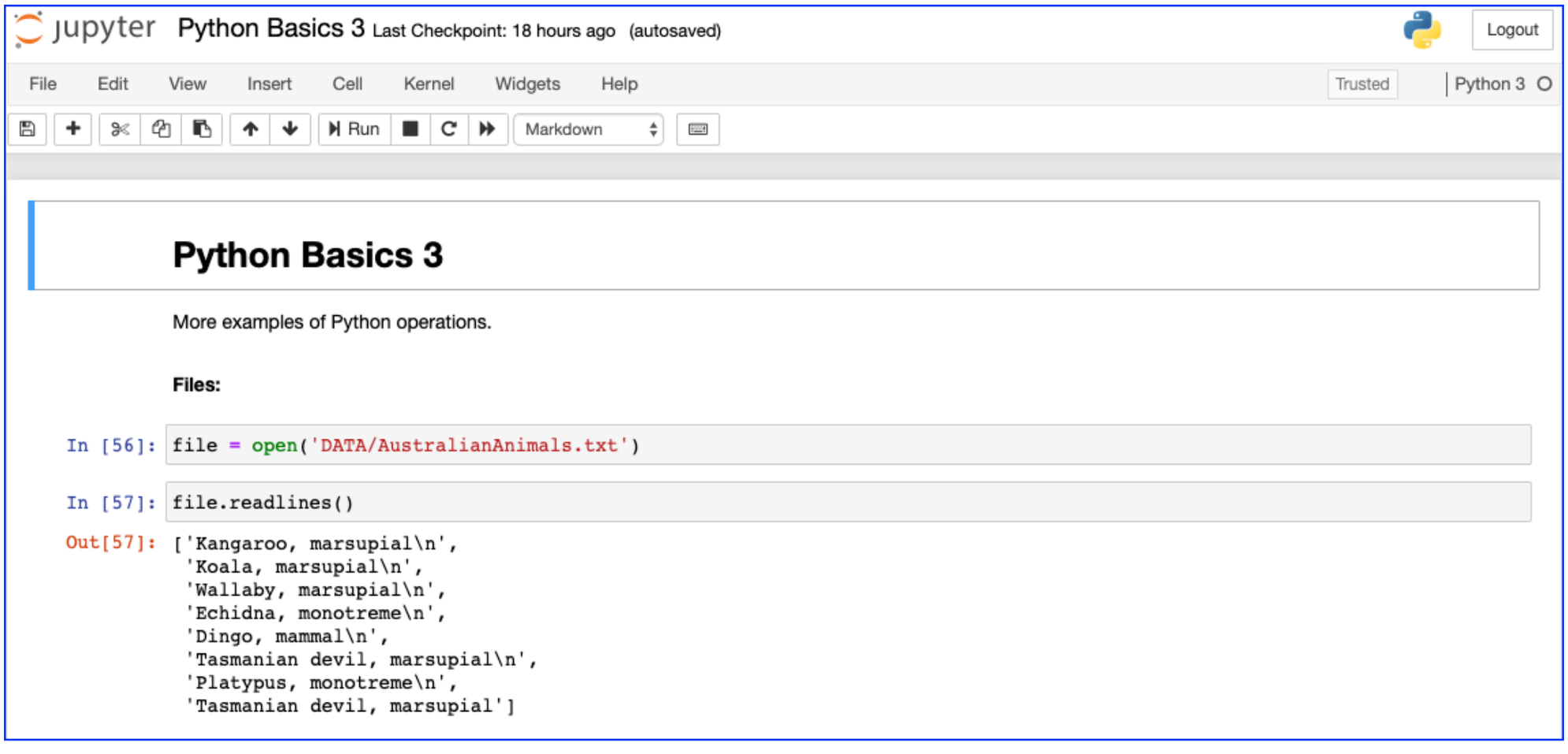


How to write Python code

Using Jupyter Notebook to run Python commands

Python Basics 3

Open the file **Python Basics 3.ipynb** by browsing for file in Jupyter Notebook



USING PANDAS LIBRARY TO EXPLORE AND VISUALISE DATA

pandas *is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language (*https://pandas.pydata.org/about.html*).*

Pandas

Examples using Jupyter Notebook to run pandas commands to Create, Clean, and Read data

Open the file **Creating, Cleaning and Reading Data.ipynb** by browsing for file in Jupyter Notebook

Creating Data

Creating sample dataframes using Pandas library

In [234]:
```
# import pandas library - we also import numpy for use in a couple examples
import pandas as pd
import numpy as np
```

Pandas is an open-source Python library used for data analysis. Here we will use it to create sample data to demostrate how it works.

More info here: https://pandas.pydata.org/

Firstly, check the version numbers of both:

In [235]:
```
pd.__version__
```

Out[235]: '0.25.1'

In [236]:
```
np.__version__
```

Out[236]: '1.17.2'

In [237]:
```
pd?
```

In [312]:
```
from IPython.display import Image
Image('Pandas_DataFrame.png')
```