Population Data Science with Python

Yaser Khorrami John Karuitha

August 30, 2023

Table of contents

Preface **3**

1 Introduction to Python for Data Science **4**
1.1 Background 4
1.2 Installing Python 4
1.2.1 Installing Python on Windows 4
1.2.2 Installing Python on Mac OS 5
1.2.3 Installing Python on Linux 5
1.3 Popular Python Text Editors and Interactive Development Environments (IDEs). 5
1.4 Setting up VS Code for Python Programming 6
1.5 Installing Python Packages 7
1.6 Loading Data into Python 7
1.7 Exploring Data in Python 8

2 Summary **11**

References **12**

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

1 Introduction to Python for Data Science

1.1 Background

In this section, we delve into the basics of Python for Data Science. Python is a simple yet powerful programming language that has utility in web development, scientific computing, data science and machine learning. For a start, there are two versions of Python; Python version 2 and Python version 3. In this course, we work exclusively with Python version 3. Moreover, our interest in this section is the use of Python for data analysis. Let us first install Python.

1.2 Installing Python

The installation of Python will differ slightly depending on the operating system; Windows, Mac, and Linux. The site https://www.python.org/downloads/ contains the Python executables for each operating system. ASt the time of writing this book, the Python version release is Python 3.11.5. However, installation procedures do not change much. The internet is full of tutorials on the installation of Python. In this book, we refer the reader to the available installation guidelines.

1.2.1 Installing Python on Windows

Microsoft has a comprehensive set of installation procedures for installing Python on Windows available on this website https://learn.microsoft.com/en-us/windows/python/beginners. Microsoft recommends the installation of Python from the Microsoft Store. We also recommend this approach because it will save you from the complications of setting the Python path. The link also contains information about the installation of VS Code, a popular text editor for writing Python code. We recommend that you also install VS Code.

If you choose to download and install Python directly from the Python Website, ensure that you set the path correctly. Specifically, when installing Python, ensure that you tick the choice `Add Python to Path` in the installation dialogue box (See Figure 1.1).

Figure 1.1: Add Python to Path

1.2.2 Installing Python on Mac OS

We refer the reader to the following website https://www.makeuseof.com/how-to-install-python-on-mac/ for instructions on installing Python on Mac OS. We spewcifically point you to the section titled "How to Install Python With the Official Installer" as it offers a simpler and direct way to install Python on Mac OS. We also recomend that the readers install VS Code by following instructions on this site https://code.visualstudio.com/docs/setup/mac.

1.2.3 Installing Python on Linux

Most linux distributions come with linux pre-installed. For instance, Ubuntu comes with the latest Linux 3 release installed. To check the version of Python on Linux, open the terminal and run the following command.

```
python3 --version
```

To install VS Code, follow the instructions on this link https://code.visualstudio.com/docs/setup/linux.

1.3 Popular Python Text Editors and Interactive Development Environments (IDEs).

There are numerous popular IDEs and text editors for use with Python. The most popular IDE is `pycharm`. Pycharm comes in two flavors, the professional edition and the community edition.

The community edition has reduced functionality compared to the professional edition.

The most popular text editor for Python is `VS Code`. VS Code is free to download and use. This is our editor of choice iin this book. Our choice of VS Code is out of our personal preference. You can follow the contents of this book while using other platforms like Sublime text, Jupyter notebooks, among others.

1.4 Setting up VS Code for Python Programming

VS Code is a text editor. To make VS Code work with Python (and other programming languages), we need to install appropriate VS Code extensions. In our case, we install the following VS Code extensions.

- Python.
- Jupyter
- Code Runner.
- Quarto
- Prettier.

Let us illustrate how to install the Python extension.

- First, open the Extensions view (Ctrl+Shift+X).
- Filter the extension list by typing 'python'.
- Click on the Python extension (Verify that it the extensions is created by Microsoft).
- Finally, Install the extension (See Figure 2 and 3 below).

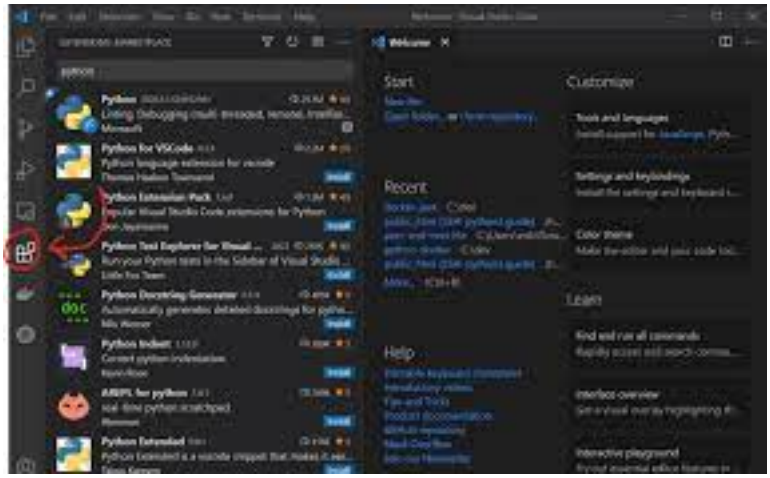

Figure 1.2: Open the extensions panel

You can follow the same procedure to install the other extensions.

Figure 1.3: Install the Python extension

1.5 Installing Python Packages

1.6 Loading Data into Python

We shall work with data from the United Nations Population Department (UNPD) to illustrate data analysis in Python. The data consists of population and life expectancy estimates and is available in the following website: https://population.un.org/wpp/Download/Standard/Most Used/.

The first step in analyzing data in Python is to load the standard libraries: pandas for importing files, matplotlib and seaborn for data visualization, and numpy for mathematical operations. When importing the libraries, it is common, though not neccesary to alias the packages (like pd for pandas and plt for matplotlib.pyplot). This convention makes it easy to reference the libraries when writing code. Not that you could use any other alias. However, in the Python community, pandas is usually aliased as pd. The same is the case for the other libraries.

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

We start by importing the data using pandas. Pandas has many handy functions for importing data in variuous formats. Given that our data is in Ms Excel format, we use the `pd.read_excel()` function to import the data.

The `pd.read_excel()` webpage details the numerous arguments that we could supply to the function. To keep things simple, we will just supply the file path. The data is in the first

sheet of the excel workbook and has column names as the first row. Hence, we stick with the default arguments; sheet_name=0, and header=0. Note that we could also supply a list of alterantive column names to the `names` parameter. For now, we leave the names parameter to the default of none.

```
population = pd.read_excel("data/WPP2022_GEN_F01_DEMOGRAPHIC_INDICATORS_COMPACT_REV1.xlsx"
```

1.7 Exploring Data in Python

The `head` method allows us to view the first 5 rows of the data table by default. In the example below, we specify that we want to display the first 3 rows instead.

```
population.head(3)
```

	index	variant	region_subregion_country_area	notes	location_code	ISO3_code	ISO2_code
0	1	Estimates	WORLD	NaN	900	NaN	NaN
1	2	Estimates	WORLD	NaN	900	NaN	NaN
2	3	Estimates	WORLD	NaN	900	NaN	NaN

We can do the same using the `tail` method to view the last few rows of the data table.

```
population.tail()
```

	index	variant	region_subregion_country_area	notes	location_code	ISO3_code	ISO2_c
20591	20592	Estimates	Wallis and Futuna Islands	2	876	WLF	WF
20592	20593	Estimates	Wallis and Futuna Islands	2	876	WLF	WF
20593	20594	Estimates	Wallis and Futuna Islands	2	876	WLF	WF
20594	20595	Estimates	Wallis and Futuna Islands	2	876	WLF	WF
20595	20596	Estimates	Wallis and Futuna Islands	2	876	WLF	WF

Let us look at the number of rows and columns of the data by calling the `shape` attribute.

```
population.shape
```

```
(20596, 65)
```

The `info()` method allows us to have an overview of the data incluyding the column names and data types.

```
population.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20596 entries, 0 to 20595
Data columns (total 65 columns):
 #   Column                                                Non-Null Count  D
---  ------                                                --------------  -
 0   index                                                 20596 non-null  i
 1   variant                                               20596 non-null  o
 2   region_subregion_country_area                         20596 non-null  o
 3   notes                                                 5475 non-null   o
 4   location_code                                         20596 non-null  i
 5   ISO3_code                                             17064 non-null  o
 6   ISO2_code                                             16992 non-null  o
 7   SDMX_code                                             20304 non-null  i
 8   type                                                  20596 non-null  o
 9   parent_code                                           20596 non-null  i
 10  year                                                  20592 non-null  i
 11  total_pop_jan1_000                                    20596 non-null  o
 12  total_pop_july1_000                                   20596 non-null  o
 13  male_pop_july1_000                                    20596 non-null  o
 14  female_pop_july1_000                                  20596 non-null  o
 15  pop_density_july1_sqkm                                20596 non-null  o
 16  pop_sex_ratio_july_males_per_100_females              20596 non-null  o
 17  median_age_july1_years                                20596 non-null  o
 18  natural_change_births_minus_deaths_000                20596 non-null  o
 19  natural_change_births_minus_deaths_per_000            20596 non-null  o
 20  pop_change_000                                        20596 non-null  o
 21  pop_growth_rate _percentage                           20596 non-null  o
 22  Pop_annual_doubling_years                             20596 non-null  o
 23  Births_000                                            20596 non-null  o
 24  births_women_15_19_000                                20596 non-null  o
 25  crude_birth_rate_per_000                              20596 non-null  o
 26  total_fertility_rate_live_births_per_woman            20596 non-null  o
 27  net_reproduction_rate_surviving_daughters_per_woman   20596 non-null  o
 28  mean_age_childbearing_years                           20596 non-null  o
 29  sex_ratio_at_birth_males_per_100_female_births        20596 non-null  o
 30  total_deaths_thousands                                20596 non-null  o
 31  male_deaths_thousands                                 20596 non-null  o
 32  female_deaths_thousands                               20596 non-null  o
 33  crude_death_rate_deaths_per_1000_population           20596 non-null  o
```

```
 34  life_expectancy_at_birth_both_sexes                                 20596 non-null  c
 35  male_life_expectancy_at_birth                                       20596 non-null  c
 36  female_life_expectancy_at_birth                                     20596 non-null  c
 37  life_expectancy_at_15_both_sexes                                    20596 non-null  c
 38  male_life_expectancy_at_15                                          20596 non-null  c
 39  female_life_expectancy_at_15                                        20596 non-null  c
 40  life_expectancy_at_65_both_sexes                                    20596 non-null  c
 41  male_life_expectancy_at_65                                          20596 non-null  c
 42  female_life_expectancy_at_65                                        20596 non-null  c
 43  life_expectancy_at_80_both_sexes                                    20596 non-null  c
 44  male_life_expectancy_at_80                                          20596 non-null  c
 45  female_life_expectancy_at_80                                        20596 non-null  c
 46  infant_deaths_under_age_1_thousands                                 20596 non-null  c
 47  infant_mortality_rate_infant_deaths_per_1000_live_births            20596 non-null  c
 48  live_birth_surviving_to_age1 _thousands                             20596 non-null  c
 49  under_five_deaths_under_age5_thousands                              20596 non-null  c
 50  under_five_mortality_deaths_under_age5_per_1000_live_births         20596 non-null  c
 51  mortality_before_age_40_both_sexes_deaths_per_1000_live_births      20596 non-null  c
 52  male_mortality_before_age_40_deaths_per_1000_male_births            20596 non-null  c
 53  female_mortality_before_age_40_deaths_per_1000_female_births        20596 non-null  c
 54  mortality_before_age_60_both_sexes_deaths_per_1000_live_births      20596 non-null  c
 55  male_mortality_before_age_60_deaths_per_1000_male_births            20596 non-null  c
 56  female_mortality_before_age_60_deaths_per_1000_female_births        20596 non-null  c
 57  mortality_age_15_50_both_sexes_deaths_under50_per_1000_alive_at_15  20596 non-null  c
 58  male_mortality_age_15_50_deaths_under50_per_1000_male_alive_at_15   20596 non-null  c
 59  female_mortality_age_15_50_deaths_under50_per_1000_female_alive_at_15  20596 non-null  c
 60  mortality_age_15_60_both_sexes_deaths_under60_per_1000_alive_at_15  20596 non-null  c
 61  male_mortality_age_15_60_deaths_under60_per_1000_male_alive_at_15   20596 non-null  c
 62  female_mortality_age_15_60_deaths_under60_per_1000_female_alive_at_15  20596 non-null  c
 63  net_migrants_000                                                    20596 non-null  c
 64  net_migration_per_1000                                              20596 non-null  c
dtypes: float64(2), int64(3), object(60)
memory usage: 10.2+ MB
```

2 Summary

In summary, this book has no content whatsoever.

References