**Data cleansing in Python (Part 1)**

This planning document is intended to support teachers who are delivering the NPA/PDA Data Science or for students who are learning independently. It also aligns with the Data Skills for Work framework.

Contents

[Version Control 1](#_Toc97034334)

[Lesson Description 2](#_Toc97034335)

[Lesson Contents 2](#_Toc97034336)

[Learning Intentions 2](#_Toc97034337)

[Success Criteria 2](#_Toc97034338)

[Knowledge Prerequisites 3](#_Toc97034339)

[Lesson Requirements 3](#_Toc97034340)

[Jupyter Notebook 4](#_Toc97034341)

[Datasets 5](#_Toc97034342)

[How you can use this lesson 6](#_Toc97034343)

[Alternative format 6](#_Toc97034344)

# Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Version number | Purpose/Change | By | Date |
| 1.0 | Published by Effini | John Bell | 10 Mar 2022 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Lesson Description

|  |  |
| --- | --- |
| **Lesson Overview** | Introduction to data cleansing activities as part of the analysis steps, including importing datasets without importing metadata; dropping unrequired rows and variables; removing duplicate rows, and renaming variables; |
| **Topic** | Data Manipulation and Data Analysis |
| **Book Chapter(s)** | Analysing data |

|  |  |
| --- | --- |
| **NPA level** | 5, 6 |
| **PDA level** | 7, 8 |
| **Data skills for work level** | Core, Analysis |

# Lesson Contents

This lesson consists of:

* A lesson plan (this document)
* A PowerPoint presentation, ‘Data Cleansing in Python (Part 1)’
* 2 Jupyter notebooks:
  + ‘data\_cleansing\_part\_1.ipynb’ (for learners)
  + ‘data\_cleansing\_with\_answers\_part\_1.ipynb’ (for teachers)

# Learning Intentions

We will be learning about data cleansing in Python,specifically,

* how to **import** a datasetwithout importing **metadata**
* what naming conventions are commonly used for variables and how to **rename variables**
* how to **drop unrequired rows** and **variables**
* how to **drop duplicates**

# Success Criteria

I can *import* a dataset without importing metadata in Python

I can *describe* different naming conventions

I can *change* the name of a variable to a chosen naming convention in Python

I can *remove* rows and variables in Python

I can *remove* duplicate rows in Python

# Knowledge Prerequisites

Learners should know:

* Python programming to at least the level defined in SQA Computer Programming Level 5 (HY2C 45)
* How to use a Jupyter notebook to write, edit and run Python code
* Data understanding is part of the analysis steps

# Lesson Requirements

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PDA** | **NPA** | **Data Skills for work** |
| **Qualification** | Yes | Yes | Yes |
| **Outcome ID(s)** | WD7.2c, WD8.3e | DS5.2c, DS5.3c,  DS6.2b | C2.1, A1.2, A2.1, A2.3 |
| **Outcome description(s)** | WD7.2c Data cleaning  WD8.3e Data cleaning | DS5.2c Describe methods of cleaning and transforming data  DS5.3c Perform routine data cleaning and structuring.  DS6.2b Perform data transformation to complete, correct and structure data | C2.1 Vocabulary used in data science and analytics  A1.2 Data quality  A2.1 Use of tools to analyse data  A2.3 Data calculation and manipulation |
| **Level** | 7, 8 | 5, 6 | Core, Analysis |
| **Software language** | Python | Python | Python |
| **Required equipment /software for student** | Lesson: PowerPoint  Python notebook: Jupyter notebook environment | Lesson: PowerPoint  Python notebook: Jupyter notebook environment | Lesson: PowerPoint  Python notebook: Jupyter notebook environment |

# Jupyter Notebook

There is a Jupyter notebook for this lesson that provides examples and programming tasks for learners, drawn from the examples in the lesson PowerPoint.

The notebook uses Python 3.x and the following packages:

* [numpy](https://numpy.org/) – for scientific computing
* [pandas](https://pandas.pydata.org/) - for data manipulation
* [s3fs](https://pypi.org/project/s3fs/) - an API to AWS S3 (Simple Storage Service), used to import datasets
* [pyjanitor](https://pyjanitor-devs.github.io/pyjanitor/) – for cleaning data

The tasks are described in the table below.

|  |  |  |
| --- | --- | --- |
| **Notebook section** | **Task** | **Description** |
| Handle Metadata | Task 1 - No Metadata for me, thanks | Import a dataset without importing the metadata contained in the csv file. |
| Rename Variables | Task 2 - Clean the names | Use the pyjanitor clean\_names() method to convert the variable names in a dataset to snake case. |
| Task 3 - Choose a better name | Choose a clear and meaningful name for a badly-named variable and rename it. |
| Task 4 - Rename the other badly-named variables | Use a data dictionary to choose clear and meaningful names for 2 badly-named variables and rename them.  Learners have the option to rename the variables one-at-a-time, or in a single line of code.  The latter requires the learner to follow online reference documentation. |
| Drop Unrequired Rows or Variables | Task 5 - Nothing useful here | Drop a row in a data frame using the pandas drop() method. |
| Task 6 - Drop multiple books | Drop two rows in a data frame using the pandas drop() method. |
| Task 7 - Dedupe the books | Drop duplicate rows in a data frame.  The learner may chose to manually do this using the drop() method or using drop\_duplicates(). |
| Task 8 - Not needed | Drop a variable in a data frame using the pandas drop() method. |
| Extension Task 1 - A good clean needed | Rename a variable, drop duplicate rows, drop empty or near-empty rows and drop empty variables in an unfamiliar dataset. |

# Datasets

The following datasets are used in this lesson.

|  |  |  |
| --- | --- | --- |
| **Dataset name** | **Description** | **Link** |
| strava\_activities | A small dataset of running and cycling activities for some [Strava](https://www.strava.com/) athletes, which requires cleaning. | <https://datasets.learn-data.science/strava_activities_small_messy.csv> |
| books | A small dataset of book review ratings from [Goodreads](https://www.goodreads.com/), which requires cleaning. | <https://datasets.learn-data.science/books_small_messy.csv> |
| employees | A small dataset of fictitious employees, which requires cleaning | <https://datasets.learn-data.science/employees_small_messy.csv> |

# How you can use this lesson

This lesson has been created by Effini in partnership with Data Education in Schools, The Data Lab and Data Skills for Work, with funding from the Scottish Government.

© 2021. This work is licensed under a [*CC BY-NC-SA 4.0 license*](https://creativecommons.org/licenses/by-nc/4.0/legalcode)*.*

A picture containing text, clipart

Description automatically generated

You are free to:

* **Share** – copy and redistribute the material in any medium or format
* **Adapt** – remix, transform and build upon the material

Under the following terms:

* **Attribution** — You must give [appropriate credit](https://creativecommons.org/licenses/by-nc/4.0/), provide a link to the license, and [indicate if changes were made](https://creativecommons.org/licenses/by-nc/4.0/). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
* **NonCommercial** — You may not use the material for [commercial purposes](https://creativecommons.org/licenses/by-nc/4.0/).
* **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](https://creativecommons.org/licenses/by-nc-sa/4.0/) as the original.

# Alternative format

**If you require this document in an alternative format, such as large print or a coloured background, please contact**

**hello@effini.com**

**or**

**4th Floor, The Bayes Centre**

**47 Potterrow**

**Edinburgh**

**EH8 9BT**