Advanced 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
Lesson Description 2
Lesson Contents 2
Learning Intentions 2
Success Criteria 2
Knowledge Prerequisites 3
Lesson Requirements 3
Jupyter Notebook 4
Datasets 4
How you can use this lesson 5
Alternative format 6

Version Control

Version number	Purpose/Change	By	Date
1.0	Published by Effini	John Bell	29 March 2022

Lesson Description

Lesson Overview	This lesson is intended to follow the **Data Cleansing in Python** lesson. Introduction to advanced data cleansing activities as part of the analysis steps, specifically converting between different data types (strings, integers and dates)
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, 'Advanced Data Cleansing in Python'
- 2 Jupyter notebooks:
    - 'advanced_data_cleansing_part_1.ipynb' (for learners)
    - 'advanced_data_cleansing_with_answers_part_1.ipynb' (for teachers)

Learning Intentions

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

- how to **convert** between different data types

Success Criteria

I can *convert* between different data types 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
- The fundamentals of data cleansing, as covered in **Data Cleansing in Python**

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 – for scientific computing
- pandas - for data manipulation
- s3fs - an API to AWS S3 (Simple Storage Service), used to import datasets
- datetime – for working with dates

The tasks are described in the table below.

Notebook section	Task	Description
Convert Data Types	Task 1 - Car parts	Convert an integer variable in a data frame to a string variable.
	Task 2 - Mars conversion	Convert a string variable in a data frame to an integer variable.
	Task 3 - Mars data dictionary	Describe what should be done to a data dictionary when a variable within the dataset it describes is converted to another data type.
	Extension Task 1 - Temperature difference	Create a new variable in a data frame using recently converted variables.
	Task 4 - Change dates on Mars	Convert a string variable in a data frame to a date variable.
	Extension Task 2 - How many days?	Calculate the difference between the maximum and minimum dates in a date variable in a data frame (in days).

Datasets

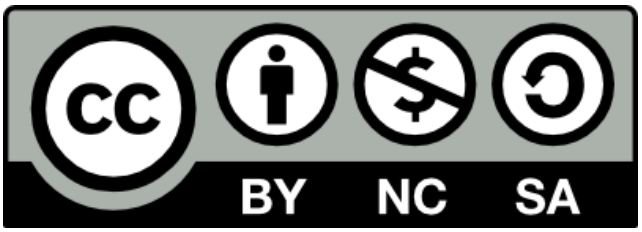
The following datasets are used in this lesson.

Dataset name	Description	Link
popular_dog_names	The most popular dog names in Anchorage, Alaska, in 2017	https://datasets.learn-data.science/popular_dog_names.csv

mars_temperature	The minimum and maximum ground temperature on Mars over a few days, as recorded by the Rover Environmental Monitoring Station (REMS) onboard the Curiosity Rover on Mars)	https://datasets.learn-data.science/mars_temperature.csv
phone_numbers	Some fictitious phone numbers	https://datasets.learn-data.science/phone_numbers.csv
car_parts	Part of the stock inventory from a car parts retailer	https://datasets.learn-data.science/car_parts.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.



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, provide a link to the license, and indicate if changes were made. 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.
- **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license 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