Data cleansing in Python (Part 2)

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

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

Lesson Description

Lesson Overview	Part 2 of an introduction to data cleansing activities as part of the analysis steps, including handling missing values and outliers.
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'
- 2 Jupyter notebooks:
    - 'data_cleansing_part_2.ipynb' (for learners)
    - 'data_cleansing_with_answers_part_2.ipynb' (for teachers)

Learning Intentions

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

- how to **handle missing** and **outlying values** that have already been identified

Success Criteria

I can *replace* all missing values with a given value in Python

I can *replace* the value of an outlier with a given value 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
- How to identify missing and outlying values using pandas
- How to filter datasets using pandas

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
- pyjanitor – for cleaning data

The tasks are described in the table below.

Notebook section	Task	Description
Handle Missing Data	Task 8 - Drop Book rows with missing data	Drop rows containing completely missing data in a data frame using the pandas dropna() method.
	Task 9 - Drop Book variables with missing data	Drop variables containing completely missing data in a data frame using the pandas dropna() method.
	Task 10 - Household energy consumption	Fill in missing values with a given value using the pandas fillna() method.
Handle Outliers	Task 11 - How to deal with outliers?	State the 3 options there are for handling an outlier.
	Task 12 - The coldest temperatures	Give the most appropriate method of handling an outlier in a dataset and give a reason for your choice.
	Task 13 - Too much energy	Give the most appropriate method of handling an outlier in a dataset and write some code to remove/replace/leave as is.

Datasets

The following datasets are used in this lesson.

Dataset name	Description	Link
books	A small dataset of book review ratings from Goodreads, which requires cleaning.	https://datasets.learn-data.science/books_small_messy.csv

cafe_tables	The number of customers sitting at each table in a café.	https://datasets.learn-data.science/cafe_tables.csv
addresses	Some UK addresses.	https://datasets.learn-data.science/addresses.csv
energy_consumption	Household energy usage (in Kilowatts per hour) over 10 days, for a single household.	https://datasets.learn-data.science/energy_consumption.csv
employees	Employee-related information.	https://datasets.learn-data.science/employees.csv
heights	The heights of some children.	https://ed-uni-data-lessons.s3.eu-west-2.amazonaws.com/data/heights.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*.

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