Advanced 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	XX March 2022

Lesson Description

Lesson Overview	This lesson is intended to follow the **Advanced Data Cleansing in Python Part 1** lesson. Introduction to advanced data cleansing activities as part of the analysis steps, including: • Fixing strings, specifically changing the case of string variables and stripping unwanted whitespace from string variables • Focusing on the causes of missing/outlying values and how these impacts on how you handle them.
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_2.ipynb’ (for learners)
  - ‘advanced_data_cleansing_with_answers_part_2.ipynb’ (for teachers)

Learning Intentions

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

- how to **fix strings**
- understand the reasons **why there may be missing values or outliers**, and how these reasons affect the ways in which we handle them

Success Criteria

I can *fix* a string in Python.

I can *describe* what might cause a missing value and how this might affect how I fix it.

I can *describe* what might cause an outlying value and how this might affect how I fix it.

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**
Fix Strings	Task 1 - Fix the dogs' names	A) Choose the correct pandas function to convert a string variable to Title case. B) convert a string variable to Title case.
	Task 2 - Import the car parts dataset without unwanted whitespace	Strip leading and trailing space in 2 string variables in a dataset when it is imported.
	Task 3 - Strip the whitespace from the car parts	Strip leading and trailing space in 2 string variables in a dataset after it has been imported.
	Extension Task 1 - Strip whitespace *and* capitalise when importing	Write a custom function to strip whitespace and capitalise and use it to perform the two conversions when importing with read_csv().
Fixing Missing Values and Outliers	Task 4 - Options for handling missing data	Name the options for handling missing data
	Task 5 - The mysterious missing death of David MacMillan	A) Identify possible reasons for missing data in a dataset B) State an appropriate way of handling the missing data

	Task 6 - World population	A) Identify and name an outlier in a dataset B) Identify possible reasons for the outlier
	Task 7 - Cafe Survey	A) Identify possible reasons for missing data in a dataset B) State an appropriate way of handling the missing data

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
phone_numbers	Some fictitious phone numbers	https://datasets.learn-data.science/phone_numbers.csv
addresses2	Some fictitious addresses	https://datasets.learn-data.science/addresses2.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:

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