Practise data cleansing in Python

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	12 April 2022

Lesson Description

Lesson Overview	This lesson is intended to follow the **Data Cleansing in Python Part 1** and **Part 2** lessons. This is a consolidation activity to give learners the chance to apply the data cleansing skills they have learned in **Data Cleansing in Python Part 1** and **Part** 2. Learners will be provided with a single dataset and cleanse it from start to finish. This dataset is the same one used in the **Practise Data Understanding** lesson. Learners should have recorded their findings during **Practise Data Understanding** lesson and will require access to them during this lesson.
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, 'Practise Data Cleansing in Python'
- 2 Jupyter notebooks:
    - 'practise_data_cleansing.ipynb' (for learners)
    - 'practise_data_cleansing_with_answers.ipynb' (for teachers)

Learning Intentions

We will be learning how to apply data cleansing techniques to **cleanse a dataset using Python**, specifically,

- how to **import** a dataset without importing **metadata**

- how to **rename variables**
- how to **drop unrequired rows** and **variables**
- how to **drop duplicates**
- how to **handle missing data** and **outliers**

Success Criteria

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

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

I can *remove* rows which contain outliers 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 cleansing follows data understanding as part of the analysis steps
- The fundamentals of data cleansing, as covered in **Data Cleansing in Python Part 1** and **Part 2**

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
Rename variables	Task 1 - Rename variables	Identify badly-named or inconsistent variable names in a dataset and rename them to something consistent and meaningful.
Drop duplicate rows	Task 2 - Drop duplicate rows	Drop any duplicate rows from the dataset.
Handle outliers and missing data	Task 3 - Drop the rows containing the outliers	Drop the rows containing the outliers using a filter.
	Task 4 - Handle missing data in 'loudness'	Drop a variable from the dataset.
	Task 5 - Handle missing data in 'decade	Replace missing values in the decade variable using a calculated field.
	Task 6 - Handle rows with no data	Remove rows where all data items are missing.

Datasets

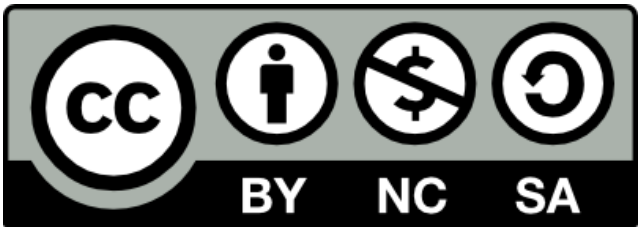
The following datasets are used in this lesson. As the lesson involves end-to-end cleansing of a dataset, to avoid learners becoming 'stuck' if they are unable to complete a step, interim datasets have been created which learners can access if needed. For example, if a learner is unable to complete the first step (removing the metadata from a data file) they can download music_metadata_removed.csv and continue to work with this dataset for the next step (renaming variables).

Dataset name	Description	Link
music_messy	This is the primary dataset for this lesson. Contains information about popular songs from 1945-	https://datasets.learn-data.science/music_messy.csv

	2019. The data is originally from Spotify. The dataset has been created by taking a clean dataset and introducing specific changes to it that require cleansing. Upon successfully applying the data cleansing steps in the Jupyter notebook, it will be returned to a clean state.	
music_metadata_removed	The music_messy dataset with metadata removed.	https://datasets.learn-data.science/music_metadata_removed.csv
music_variables_renamed	The music_metadata_removed dataset with variables renamed.	https://datasets.learn-data.science/music_variables_renamed.csv
music_deduped	The music_variables_renamed dataset with duplicate rows removed.	https://datasets.learn-data.science/music_deduped.csv
music_outliers_removed	The music_deduped dataset with outliers removed.	https://datasets.learn-data.science/music_outliers_removed.csv
music_variables_dropped_or_replaced	The music_outliers_removed dataset with one variable dropped and another with missing values replaced.	https://datasets.learn-data.science/music_variables_dropped_or_replaced.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