Dataset understanding 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

Version number	Purpose/Change	By	Date
1.0	Published by Effini	John Bell, Effini	11 Feb 2022

Lesson Description

Lesson Overview	The following aspects of the data understanding step in the analysis process: • metadata and data dictionaries • the size, shape and format of a dataset • the data types of variables in a dataset
Topic	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, 'Dataset Understanding in Python (Part 1)'
- Jupyter notebooks:
    - 'understanding_datasets_with_answers_part_1.ipynb' (for teachers), and
    - 'understanding_datasets_part_1.ipynb' (for learners)
- Datasets used in the Jupyter notebook: the datasets are stored online and imported by the Jupyter notebook.

The Jupyter notebook for teachers contains answers to the tasks set for learners.

Learning Intentions

We will be learning about the data understanding part of the analysis process, specifically,

- what is **metadata** and the importance of a **data dictionary**
- how to find the **shape**, **size** and **format** of datasets, using Python
- how to find the **data types** of variables in a dataset, using Python

Success Criteria

I can *describe* what metadata is and how it can be used.

I can *describe* what is a data dictionary is and how it can be used.

I can *describe* the shape, size and format of datasets, using Python.

I can *state* the data types used in a dataset, using Python.

Knowledge Prerequisites

Learners should know:

- Data is held in structured data frames
- Python is a programming language that can be used for data analysis
- How to use a Jupyter notebook to write, edit and run Python code

- Data understanding is part of the analysis process

If you wish learners to undertake the section on **Format of a dataset** (i.e. wide vs long formats), learners should have undertaken the **Reshaping Datasets** lesson first, otherwise this section should be skipped.

Lesson Requirements

	PDA	NPA	Data Skills for work
Qualification	Yes	Yes	Yes
Outcome ID(s)	CD7.1c, CD7.1f, WD8.1e, WD8.1f	DC5.2b, DC6.2b	A1.2, A1.3, C2.1
Outcome description(s)	CD7.1c Types of data CD7.1f Data quality WD8.1e Data quality WD8.1f Stages in the data analysis process	DC5.2b Explain how data can be analysed, DC6.2b Explain how data can be analysed	A1.2 Data quality A1.3 Interpretation and insight C2.1 Vocabulary used in data science and analytics
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

Python Notebook

There is a Python 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
-

The notebooks can be used with any Jupyter notebook environment. The tasks are described in the table below.

Notebook section	Task	Description
Use Metadata	Task 1 - IMDd Metadata	Use the IMDb website to provide examples of metadata that has been captured for all films.
	Task 2 - Music Metadata	Find some metadata for your favourite song.
Use a Data Dictionary	Task 3 - What are you looking at?	Use a data dictionary to learn about the variables in a dataset.
Identify the Shape, Size and Data Types	Task 4 – Number of columns	Calculate the number of columns in a dataset using the pandas **shape** property.
	Task 5 – Check it	Verify that the number returned by the pandas **size** property is 'number of rows' x 'number of columns'
	Task 6 - Size and shape	Find out the size and shape of a dataset.
	Task 7 - How does that compare?	Find out the data types in a dataset and compare this to the information provided in the data dictionary for the dataset.
	Task 8 – Wide or long?	Identify whether 2 datasets are in a wide or long format.

Datasets

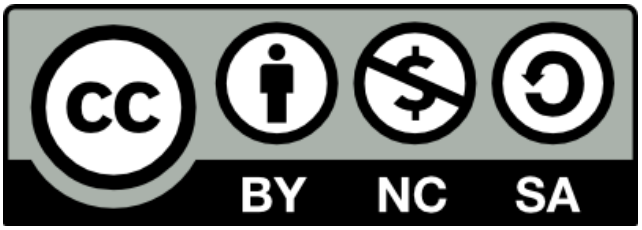
The following datasets are used in this lesson.

Dataset name	Description	Link

gold_yearly	The prices of gold from 1969 to 2021	https://datasets.learn-data.science/gold_yearly.csv
gbbo_ingredients	Ingredients used by contestants in The Great British Bakeoff, Series 12	https://datasets.learn-data.science/gbbo_ingredients.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