Practise data understanding 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 Understanding in Python Part 1** and **Part 2** lessons. This is a consolidation activity to give learners the chance to apply the data understanding skills they have learned in **Data Understanding in Python Part 1** and **Part** 2. Learners will be provided with a single, unfamiliar dataset and undertake a set of activities to get a better understanding of it. Note that learners will use what they find out about this dataset in a later lesson, **Practise Data Cleansing in Python**, where they will take what they have learned and apply it to the task of cleansing the same dataset as used in this lesson. It is recommended that learners record their findings from this lesson so that they can access this during **Practise Data Cleansing in Python**.
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 Understanding in Python'
- 2 Jupyter notebooks:
  - 'practise_data_understanding.ipynb' (for learners)
  - 'practise_data_understanding _with_answers.ipynb' (for teachers)

Learning Intentions

We will be learning how to apply data understanding techniques to **understand an unfamiliar dataset using Python**, specifically,

- how to **import** a dataset without importing **metadata**
- how to use a **data dictionary** to find out about a dataset
- 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
- how to **identify outliers** and **missing values** in Python

Success Criteria

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

I can *use* a data dictionary to find out about a dataset

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

I can *find* the data types of variables in a dataset, using Python

I can *identify* outliers and missing values 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 understanding, as covered in **Data Understanding in Python**

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

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

The tasks are described in the table below.

Notebook section	Task	Description
Remove the metadata	Task 1 - Import the dataset	Import a dataset that contains metadata, without importing the metadata.
Use the metadata	Task 2 - Use the data dictionary	Use a data dictionary to find out some basic information about an unfamiliar dataset.
Find the shape, size and data types of the dataset	Task 3 - Find the shape, size and data types of the dataset	Find out some basic attributes of an unfamiliar dataset.
Identify the format of the dataset	Task 4 - Identify the format of the dataset	Through visual inspection, determine if a dataset is wide or long.
Identify outliers and missing data	Task 5 - Spot the outliers using describe()	Identify outliers in a dataset.
	Task 6 - Spot the outliers using sort_values()	Identify outliers in a dataset.
	Task 7 - Find the rows where *all* the values are missing	Find the rows where *all* the values are missing.

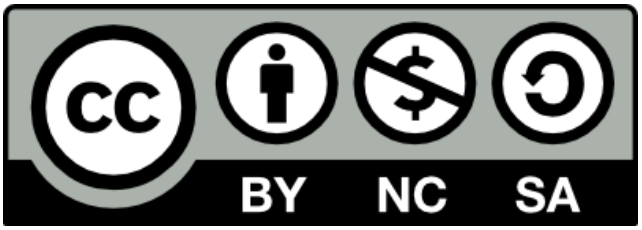
Datasets

The following dataset is used in this lesson.

Dataset name	Description	Link
music_messy	This is the only dataset used in this lesson. Contains information about popular songs from 1945-2019. 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.	https://datasets.learn-data.science/music_messy.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.



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