Dataset understanding 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, Effini	11 Feb 2022

Lesson Description

Lesson Overview	Identification of outliers and missing data in the data understanding step in the analysis process.
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 2)'
- Jupyter notebooks:
    - 'understanding_datasets_with_answers_part_2.ipynb' (for teachers), and
    - 'understanding_datasets_part_2.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,

- how to **identify outliers** and **missing values** in Python

Success Criteria

I can *identify* outliers and missing values in a dataset in 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

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, A3.1, 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 A3.1 Visualisation of data to provide 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
Identify Outliers	Task 1 - Gold!	Use the describe() or sort_values() pandas methods to find the minimum and maximum values for a variable in a dataset.
	Task 2 - How long have you worked here?	Find the outliers for a variable in a dataset.
Identify Missing Data	Task 3 - What's Missing?	Convert all missing values (which are encoded using different 'missing data' identifiers) into NaN when importing the dataset.
	Task 4 - Hey Dude, Where's My Distributor?	Identify missing data in a single variable by looking for unique values.
	Task 5 - Finding the Missing	Convert all missing values (which are encoded using different 'missing data' identifiers) into NaN when importing the dataset.

Datasets

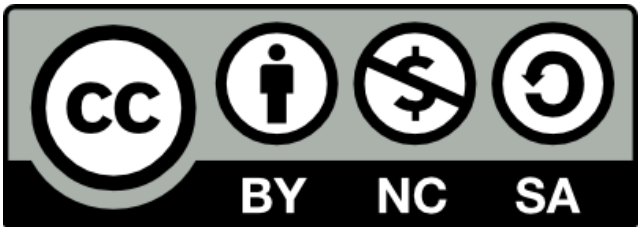
The following datasets are used in this lesson.

Dataset name	Description	Link
gold_yearly	The prices of gold from 1969 to 2021	https://datasets.learn-data.science/gold_yearly.csv
employees	Information about a fictitious set of employees in a company.	https://datasets.learn-data.science/employees.csv
empty_values2	Contains only missing data but which uses a variety of identifiers to indicate the missing values.	https://datasets.learn-data.science/empty_values2.csv

speed_skating_winter_olympics_2018_small	A small set of results from the 2018 Winter Olympics speed skating events, with some values encoded as missing.	https://datasets.learn-data.science/speed_skating_winter_olympics_2018_small.csv
highest_grossing_usa_movies_1995_2021_missing_values	the film which took the most money at the box-office in the USA from each year from 1995 to 2021, with some values encoded as missing.	https://datasets.learn-data.science/highest_grossing_usa_movies_1995_2021_missing_values.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