Manipulating dataset rows 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

Lesson Description

Lesson Overview	Subsetting Filtering Sorting Deduplicating
Topic	Data manipulation
Book Chapter(s)	“Data Transformation and Manipulation”

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, ‘Manipulating dataset rows in Python’
- Jupyter notebooks:
    - ‘data_manipulation_of_rows_with_answers.ipynb’ (for teachers), and
    - ‘data_manipulation_of_rows.ipynb’ (for learners)
- Datasets used in the Jupyter notebook: the datasets are stored in ‘the cloud’ and imported by the Jupyter notebook.

Learning Intention

We will be learning how to manipulate data in Python, specifically to be able to:

- **filter and sort** rows
- **subset** data to select the parts of the data you are interested in
- **remove duplicates** from the data

Success Criteria

I can *describe* what is means to sort, filter, subset and remove duplicates from a dataset.

I can *manipulate* data by sorting, filtering, subsetting and removing duplicates 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
- How to open a Jupyter notebook

Lesson Requirements

	PDA	NPA	Data Skills for work
Qualification	Yes	Yes	Yes
Outcome ID(s)	WD8.3b, WD8.3c, CD8.1g, WD7.2a, WD7.2b, CD7.3a	DS5.2c, DS5.3c, DS6.2b, DS6.3c	C2.1, A1.2, A2.3
Outcome description(s)	WD8.3b Types of data transformation WD8.3c Transformations CD8.1g Preparing data for visualisation WD7.2a Types of data transformation WD7.2b Common transformations including filtering, sorting CD7.3a Preparing data for visualisation *N.B. out of scope of this lesson,* *"WD8.3c … including joins"* *"WD7.2b ….combining, separating and grouping"*	DS5.2c Describe methods of cleaning and transforming data DS5.3c Perform routine data cleaning and structuring. DS6.2b Explain techniques for data capture, cleaning and transformation including data modelling DS6.3c Perform data transformation to complete, correct and structure data	C2.1 Vocabulary used in data science and analytics A1.2 Data quality A2.3 Data calculation and manipulation *N.B. out of scope of this lesson "A1.1….quantitative and qualitative"*
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:

- 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
Sort rows	Task 1- Sorting Mountains	Sort rows in a data frame alphabetically in ascending order
	Extension Task 1 - The Dawn's Early Light	Sort rows in a data frame numerically in ascending order
Filter rows	Task 2 - Later Dawn	Filter the rows in a data frame using a ‘less than’ (<) operator on a named column, either in 3 steps (with step-by-step guidance) or a single step
	Task 3 - Blowing in the Wind	Filter the rows in a data frame using a ‘greater than’ (>) operator on a named column, either in 3 steps (with step-by-step guidance) or a single step
	Extension Task 2 - Tall Peaks	Filter the rows in a data frame where you need to select the correct column to filter on and the correct operator to use
	Task 4 - Hats	Filter the rows in a data frame using a ‘equality’ (==) operator on a named column, either in 3 steps (with step-by-step guidance) or a single step
	Task 5 - Something Brighter	Filter the rows in a data frame using a ‘inequality’ (!=) operator on a named column, either in 3 steps (with step-by-step guidance) or a single step
	Extension Task 3 - Small or Medium Please	Filter the rows in a data frame where you need to select the correct column to filter on and the correct operator to use
Subsetting	Task 6 – Low Winds	Subset the rows in a data frame where the columns to select are

		specified and the column to be used to filter on is specified.
	Task 7 – Sunny Places	Subset the rows in a data frame where the columns to select are specified and the column to be used to filter on is specified.
	Extension Task 4 - Anywhere but Rockcliffe!	Subset the rows in a data frame where you need to choose the correct columns to select and the correct column to filter on
Remove duplicates	Extension Task 5 - What Will Happen?	Hypothesise about what executing a function to deduplicate the rows in a data frame will do when the data frame contains no duplicates, and test your hypothesis by executing the function and comparing the original and new data frames

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.

© 2021. This work is licensed under a *CC BY-NC-SA 4.0 license*.

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.