Summarising datasets 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

Lesson Description

Lesson Overview	Summarising complete datasets and selected variables from a dataset, using summary calculations such as the total, count, min/max and mean.
Topic	Data manipulation/ Data analysis
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, 'Summarising datasets in Python Part 1'
- Jupyter notebooks:
    - 'summarising_datasets_with_answers_part_1.ipynb' (for teachers), and
    - 'summarising_datasets_part_1.ipynb' (for learners)
- Datasets used in the Jupyter notebooks: the datasets are stored online and imported by the Jupyter notebooks.

Learning Intentions

We will be learning to summarise datasets in Python, specifically to:

- summarise complete datasets
- perform summary calculations for single variables, such as the total, count, min/max and average values
- perform summary calculations for multiple variables

Success Criteria

I can *describe* how to group rows of data based on logical criteria.

I can *group* and *summarise* rows of data 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 to write, edit and run Python code

Lesson Requirements

	PDA	NPA	Data Skills for work
Qualification	Yes	Yes	Yes
Outcome ID(s)	WD8.3c, CD8.1g, WD7.3d, WD7.2a, WD7.2b, CD7.3a	DS5.3d, DS6.3d	C2.1, A2.1, A2.3
Outcome description(s)	WD8.3c Transformations CD8.1g Preparing data for visualisation WD7.3d Data aggregation 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"*	DS5.3d Perform analyses including […] group and summarise, DS6.3d Perform descriptive and predictive analyses on the data.	C2.1 Vocabulary used in data science and analytics 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:

- 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
Summarise a complete data frame	Task 1 - In a Galaxy Far Far Away	Summarising a data frame using the pandas describe() method.
Summarise a single variable	Task 2 - The Colour of Aliens	Producing some summary calculations for a single variable using the pandas describe() method.
	Task 3 - The Tallest Alien	Calculating the maximum value for a numeric variable.
	Task 4 - Mean Mass	Calculating the mean value for a numeric variable.
	Task 5 - All the Eye Colours	Calculating the count of values for a text variable.
	Task 6 - How Many Different Species?	Calculating the count of unique values for a text variable.
Summarise multiple variables	Task 7 - How Small and Light	Calculating the minimum values for two named numeric variables.
	Task 8 - How Many Unique Values?	Calculating the count of unique values for all of the text variables in a data frame.

Datasets

The following datasets are used in this lesson.

Dataset name	Description	Link
Fruit & veg	The sales of fruit and vegetables in a shop.	https://datasets.learn-data.science/fruit_and_veg.csv
Star Wars characters	Star Wars character data (from https://swapi.dev/).	https://datasets.learn-data.science/star_wars_characters.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.