**Practise Combining Datasets**

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 1](#_Toc132350193)

[Lesson Description 2](#_Toc132350194)

[Lesson Contents 2](#_Toc132350195)

[Learning Intentions 3](#_Toc132350196)

[Success Criteria 3](#_Toc132350197)

[Knowledge Prerequisites 3](#_Toc132350198)

[Lesson Requirements 4](#_Toc132350199)

[Task-types 5](#_Toc132350200)

[Worksheet 6](#_Toc132350201)

[How you can use this lesson 8](#_Toc132350202)

[Alternative format 9](#_Toc132350203)

# Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| Version number | Purpose/Change | By | Date |
| 1.0 | Published by effini | John Bell | 24 May 2023 |
| 1.1 | * Changed term “Outer (full) join” to “outer join” * Changed term “key column” to “key” * Made data definitions consistent with those used in Combining Datasets lesson | John Bell | 10 Aug 2023 |

# Lesson Description

|  |  |
| --- | --- |
| **Lesson Overview** | Practise joining and appending simple datasets in Python. |
| **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, ‘Practise Combining Datasets in Python’
* Jupyter notebooks:
  + ‘practise\_combining\_datasets\_with\_answers.ipynb’ (for teachers), and
  + ‘practise\_combining\_datasets.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 about **how to combine datasets in Python**, specifically

* how to append rows to a dataset, and
* how to join columns to a dataset

# Success Criteria

I can *use* Python to append rows to a simple dataset.

I can *use* Python to left join simple datasets.

I can *use* Python to right join simple datasets.

I can *use* Python to inner join simple datasets.

I can *use* Python to outer join simple datasets.

# Knowledge Prerequisites

Learners should know:

* what data is
* data can be transformed into valuable information
* data can be used to solve problems and find answers to questions
* data can be stored in different types
* how rows and columns in a dataset can manipulated
* the theory behind combining datasets

Learners should complete the **Combining Datasets** lesson before completing this lesson.

# Lesson Requirements

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PDA** | **NPA** | **Data Skills for work** |
| **Qualification** | Yes | Yes | Yes |
| **Outcome ID(s)** | WD8.3c, CD8.1g, WD7.2a, WD7.2b, CD7.3a | DS5.3c, DS5.3d, DS6.2b, DS6.3c | A2.1, A2.3, C2.1, |
| **Outcome description(s)** | WD8.3c - Transformations including joins,  CD8.1g - Preparing data for visualisation,  WD7.2a - Types of data transformation,  WD7.2b - Common transformations including filtering, sorting, combining, separating and grouping,  CD7.3a - Preparing data for visualisation | DS5.3c - Perform routine data cleaning and structuring,  DS5.3d - Perform analyses including query, sort, filter, consolidate, group and summarise,  DS6.2b - Explain techniques for data capture, cleaning and transformation including data modelling,  DS6.3c - Perform data transformation to complete, correct and structure data. | A2.1 - Use of tools to analyse data,  A2.3 - Data calculation and manipulation,  C2.1 - Vocabulary used in data science and analytics, |
| **Level** | 7/8 | 5/6 | Core, Analysis |
| **Software language** | N/A | N/A | N/A |
| **Required equipment /software for student** | Lesson: PowerPoint/PDF  Python notebook: Jupyter notebook environment | Lesson: PowerPoint/PDF  Python notebook: Jupyter notebook environment | Lesson: PowerPoint/PDF  Python notebook: Jupyter notebook environment |

# Task-types

In the worksheet for this lesson, there are up to 6 task-types to that become increasingly challenging to support the students learning. Based on the student’s previous knowledge it is possible to select the task-types that are relevant to their stage.

|  |  |
| --- | --- |
| **Task-type** | **Description** |
| **1. Recall** | To be able to recognise definitions or procedures. |
| **2. Define** | To be able to define definitions or procedures. |
| **3. Rephrase** | To be able to use their own words to describe definitions or procedures. |
| **4. Apply** | To be able to apply definitions or procedures to problem-solving activities. |
| **5. Create** | To be able to apply definitions or procedures and create their own solutions to a defined problem. |
| **6. Active** | Using knowledge from the lesson which they apply to scenarios they have researched/designed themselves. |

# 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](https://pandas.pydata.org/) - for data manipulation
* [s3fs](https://pypi.org/project/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** |
| Appending Rows | Task 1 - Fast food | Appending 2 datasets |
| Task 2 - Make sweet music? | Assessing a dataset can be appended to another. |
| Joining – Left Join | Task 3 - We're getting the band back together | Left joining 2 datasets |
| Task 4 - Day trippers | Left joining 2 datasets |
| Joining – Right Join | Task 5 - Let there be rock | Right joining 2 datasets |
| Joining – Inner Join | Task 6 - Everybody needs to be in a band | Inner joining 2 datasets |
| Task 7 - What cars have been rented? | Inner joining 2 datasets |
| Joining – Outer Join | Task 8 - Rentals and Customers | Outer joining 2 datasets |
| Task 9 - Curling | Outer joining 2 datasets |
| Recap on Joining | Task 10 – How many rows? | Working out the number of rows in the final data set after each of the join types |

# Datasets

The following datasets are used in this lesson.

|  |  |  |
| --- | --- | --- |
| **Dataset name** | **Description** | **Link** |
| curling\_men | Gold medal winners of the men's Winter Olympics curling event | <https://datasets.learn-data.science/winter_olympics_curling_men.csv> |
| curling\_women | Gold medal winners of the women's Winter Olympics curling event | <https://datasets.learn-data.science/winter_olympics_curling_women.csv> |
| mcdonalds | Nutritional information about a selection of products sold by McDonalds | <https://datasets.learn-data.science/fast_food_nutrition_mcdonalds.csv> |
| subway | Nutritional information about a selection of products sold by Subway | <https://datasets.learn-data.science/fast_food_nutrition_subway.csv> |
| alc\_cons\_13 | The percentage of 13 year-olds in Scotland who have ever had an alcoholic drink (2006-2015) | <https://datasets.learn-data.science/perc_alc_consumption_13_year_olds.csv> |
| alc\_cons\_15 | The percentage of 15 year-olds in Scotland who have ever had an alcoholic drink (2006-2015) | <https://datasets.learn-data.science/perc_alc_consumption_15_year_olds.csv> |
| musicians | Some musicians and the instruments they play | <https://datasets.learn-data.science/musicians.csv> |
| bands | Musicians and the band they are in | <https://datasets.learn-data.science/bands.csv> |
| albums | Bands and some of their albums | <https://datasets.learn-data.science/albums.csv> |
| songs | Albums and the songs on them | <https://datasets.learn-data.science/songs.csv> |
| dog\_heights | The maximum heights of selected dog breeds | <https://datasets.learn-data.science/dog_breeds_simple_max_heights.csv> |
| dog\_good\_with | Whether selected dog breeds are good with children and other dogs | <https://datasets.learn-data.science/dog_breeds_simple_good_with.csv> |
| co2\_emmissions\_by\_country | CO2 emissions by countries worldwide between 1960 and 2019 | <https://datasets.learn-data.science/co2_emmissions_by_country.csv> |
| iso\_country\_codes | 2-letter and 3-letter country codes and country names from the [International Organization for Standardization (ISO)](https://www.iso.org/) | <https://datasets.learn-data.science/ISO-country-codes.csv> |
| visit\_scotland\_visits\_2019 | The number of tourist day visits to various types of location in Scotland | <https://datasets.learn-data.science/visit_scotland_visits_2019.csv> |
| visit\_scotland\_expenditure\_2019 | The expenditure of tourists on day visits to various types of location in Scotland | <https://datasets.learn-data.science/visit_scotland_expenditure_2019.csv> |
| cars | A dataset of fictitious cars, their registration numbers, makes and models | <https://datasets.learn-data.science/cars.csv> |
| customers | The customers of a car and van rental company | <https://datasets.learn-data.science/customers.csv> |
| car\_rentals | Customers who have rented a car, what car they rented, and when | <https://datasets.learn-data.science/car_rentals.csv> |
| top\_unis | The Top 25 universities in the world, according to [The Times Higher Education](https://www.timeshighereducation.com/world-university-rankings/2023/world-ranking) | <https://datasets.learn-data.science/top_25_universities_2023.csv> |
| unis | A reference dataset of all universities in the world | <https://datasets.learn-data.science/worldwide_universities.csv> |
| country\_codes | A reference dataset of countries and their [ISO country codes](https://www.iso.org/iso-3166-country-codes.html) | <https://datasets.learn-data.science/ISO-country-codes-de-duped.csv> |

# How you can use this lesson

This lesson has been created by effini in partnership with The Data Lab.

© 2023. This work is licensed under a [*CC BY-NC-SA 4.0 license*](https://creativecommons.org/licenses/by-nc/4.0/legalcode)*.*

A picture containing text, clipart

Description automatically generated

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](https://creativecommons.org/licenses/by-nc/4.0/), provide a link to the license, and [indicate if changes were made](https://creativecommons.org/licenses/by-nc/4.0/). 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](https://creativecommons.org/licenses/by-nc/4.0/).
* **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](https://creativecommons.org/licenses/by-nc-sa/4.0/) 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**