# Creating new variables by extracting and combining 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** | Creating new variables by extracting or combining data from other variables |
| **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, ‘Creating new variables by extracting and combining in Python’
- Jupyter notebooks:
    - ‘creating_variables_by_extracting_or_combining_with_answers.ipynb’ (for teachers), and
    - ‘creating_variables_by_extracting_or_combining.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 how to create new variables in Python, specifically to,

- understand what it means to extract data to create a new variable
- create simple new variables by extracting data in Python
- understand what it means to combine data to create a new variable
- create simple new variables by combining data in Python.

## Success Criteria

I can *describe* how to create a new variable by extracting data.

I can *create* new variables in Python by extracting data.

I can *describe* how to create a new variable by combining data.

I can *create* new variables in Python by combining data.

## 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)** | WD8.3b, 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<br>WD8.3c Transformations<br>CD8.1g Preparing data for visualisation<br>WD7.2a Types of data transformation<br>WD7.2b Common transformations including filtering, sorting<br>CD7.3a Preparing data for visualisation<br>*N.B. out of scope of this lesson,*<br>*“WD8.3c … including joins”*<br>*“WD7.2b ….combining, separating and grouping”* | DS5.2c Describe methods of cleaning and transforming data<br>DS5.3c Perform routine data cleaning and structuring.<br>DS6.2b Explain techniques for data capture, cleaning and transformation including data modelling<br>DS6.3c Perform data transformation to complete, correct and structure data<br>*N.B. out of scope of this lesson,*<br>*“DS5.3d …including sort, filter…, group and summarise.”* | C2.1 Vocabulary used in data science and analytics<br>A1.2 Data quality<br>A2.3 Data calculation and manipulation<br>*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<br>Python notebook: Jupyter notebook environment | Lesson: PowerPoint<br>Python notebook: Jupyter notebook environment | Lesson: PowerPoint<br>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 |
|---|---|---|
| Create a New Variable By Extracting Data From an Existing Variable | Task 1 - Cyclists' Initials | Creating a new variable by slicing from another text variable. |
| | Task 2 - Athletes' First Names | Creating a new variable by extracting from another text variable using a regular expression. |
| | Extension Task 1 - Athletes' Full Names | Creating a new variable by extracting from another text variable using two regular expressions. |
| | Task 3 - Birthday Cards | Creating a new variable by extracting a month from a datetime variable. |
| Create a New Variable by Combining Data Items | Task 4 - Name and Team | Creating a new variable by extracting strings from two text variables and concatenating them. |
| | Extension Task 2 – Name and Wins | Creating a new variable by extracting values from a numeric and text variable, converting the numeric variable to text and concatenating them. |
| | Extension Task 3 – Name, Team and Wins | Creating a new variable by extracting values from a numeric and text variable, converting the numeric |

| | | |
|---|---|---|
| | | variable to text and concatenating them. |

## Datasets

The following datasets are used in this lesson.

| Dataset name | Description | Link |
|---|---|---|
| Archery | The scores in an archery competition | https://datasets.learn-data.science/archery.csv |
| CHI | The Community Health Index number of some fictional patients in Scotland. | https://datasets.learn-data.science/chi.csv |
| Fictional characters | The names and addresses of fictional characters from books and films. | https://datasets.learn-data.science/fictional_characters.csv |
| Athletes' birth dates | The dates of birth of some famous athletes. | https://datasets.learn-data.science/athletes_birthdates.csv |
| Tour de France winners | 20th century winners of the Tour de France. | https://datasets.learn-data.science/tour_de_france_winners.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.