# Unit 4: Data Cleaning and Transformation

## **Overview**

In this unit, I explored what it takes to prepare raw data for meaningful analysis. We broke down each stage of the **data management pipeline**, understanding how data flows from initial collection to the final point of analysis and visualisation. The focus was on how to clean, structure, and transform data in a way that ensures reliability and usability across systems.

## **Key Learning Areas**

1. **Data Cleaning Techniques**:

We looked at how to deal with inconsistencies, outliers, missing values, and formatting issues that can distort insights. As Kazil & Jarmul (2016) suggest, this includes:

* Identifying bad data and outliers
* Vetting the reliability of sources
* Matching inconsistent strings
* Spotting patterns or anomalies that affect structure

1. **The Full Data Management Pipeline (EMC, 2015):**

I learned the step-by-step process involved in managing data effectively:

* **Capturing raw data**
* **Data cleaning**
* **Data integration**
* **Database design**
* **Data analysis**
* **Data presentation and visualisation**

Each step feeds into the next, and errors early in the pipeline can carry through and affect final outcomes, making the cleaning and integration stages especially critical.

1. **Automating Data Processes**:

We also covered what it takes to automate data workflows—using tools, libraries, and consistent structure to reduce manual handling and improve efficiency.

## **Additional Concepts Covered**

1. **Data models vs Data Architecture**:

I learned to distinguish between data models (which define relationships between data elements) and data architecture (which focuses on how data is captured, organised, and structured across a system).

1. **Python Tools in Practice**:

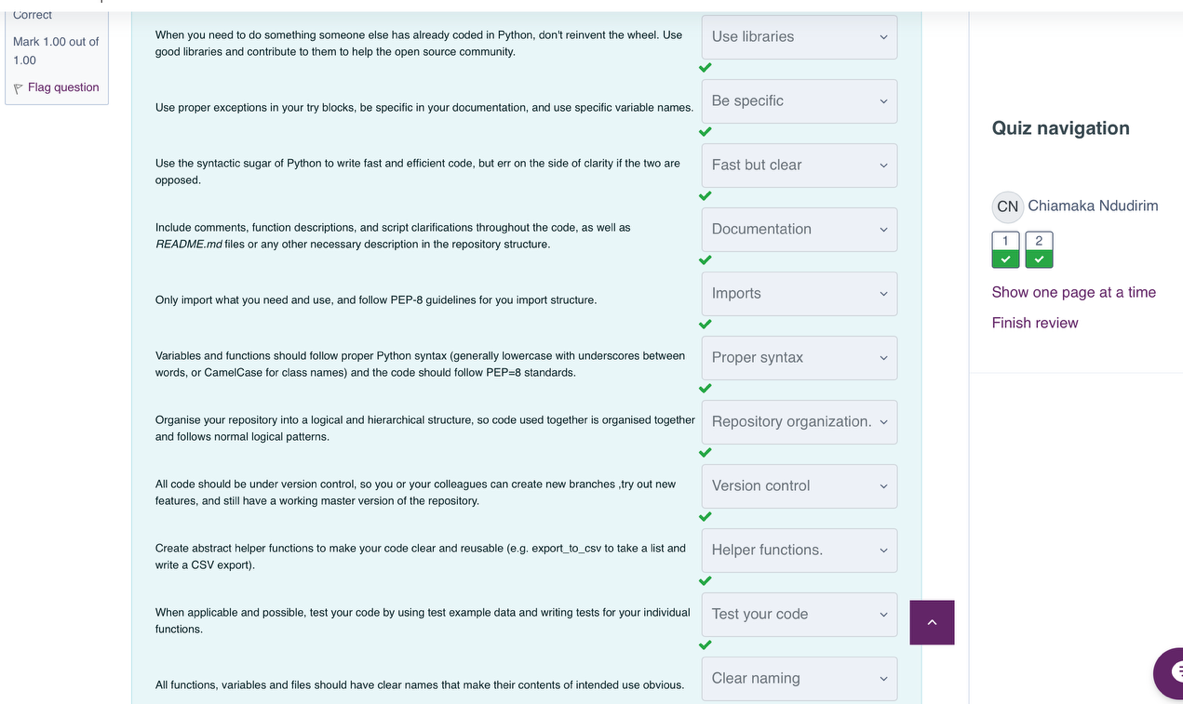
As part of our formative activity, I practiced matching key Python concepts and libraries—such as Pandas and NumPy—to real data tasks like cleaning, transformation, and validation.

## Formative Activity

We had a formative activity to complete a Data Management pipeline test. Instructions were to match python concepts or libraries to their purposes and to match python practices with the suitable description, as seen below.

A screenshot of a computer

AI-generated content may be incorrect.



## **Personal Reflection**

This unit helped me connect the dots between raw data and final analysis. It became clear that the way data is prepared behind the scenes is just as important as the analysis itself. With a better understanding of the data pipeline and how to apply structured cleaning techniques, I feel more confident in managing real-world data problems—especially when scaling or automating processes.

## Core Readings

* McKinney, W. (2022) Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter. 3rd edn. Sebastopol, California: O'Reilly.
  + Chapter 7
* Huxley et al. (2020) Data Cleaning. Sage Foundation.