# Data Engineer Certification - Practical Exam - Supplement Experiments

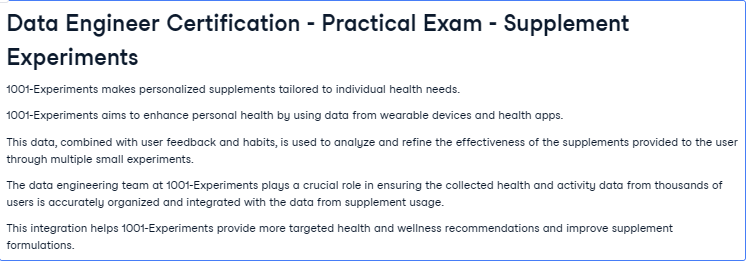
1001-Experiments makes personalized supplements tailored to individual health needs.

1001-Experiments aims to enhance personal health by using data from wearable devices and health apps.

This data, combined with user feedback and habits, is used to analyze and refine the effectiveness of the supplements provided to the user through multiple small experiments.

The data engineering team at 1001-Experiments plays a crucial role in ensuring the collected health and activity data from thousands of users is accurately organized and integrated with the data from supplement usage.

This integration helps 1001-Experiments provide more targeted health and wellness recommendations and improve supplement formulations.



## Task

1001-Experiments currently has the following four datasets with four months of data:

- "user\_health\_data.csv" which logs daily health metrics, habits and data from wearable devices,

- "supplement\_usage.csv" which records details on supplement intake per user,

- "experiments.csv" which contains metadata on experiments, and

- "user\_profiles.csv" which contains demographic and contact information of the users.

Each dataset contains unique identifiers for users and/or their supplement regimen.

The developers and data scientsits currently manage code that cross-references all of these data sources separately, which is cumbersome and error-prone.

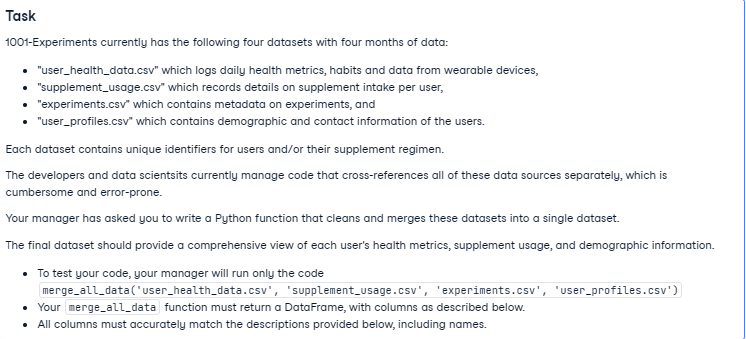
Your manager has asked you to write a Python function that cleans and merges these datasets into a single dataset.

The final dataset should provide a comprehensive view of each user's health metrics, supplement usage, and demographic information.

- To test your code, your manager will run only the code `merge\_all\_data('user\_health\_data.csv', 'supplement\_usage.csv', 'experiments.csv', 'user\_profiles.csv')`

- Your `merge\_all\_data` function must return a DataFrame, with columns as described below.

- All columns must accurately match the descriptions provided below, including names.



## Data

The provided data is structured as follows:

![database schema](schema.png)

The function you write should return data as described below.

There should be a unique row for each daily entry combining health metrics and supplement usage.

Where missing values are permitted, they should be in the default Python format unless stated otherwise.

| Column Name | Description |

|--------------------|-------------|

| user\_id | Unique identifier for each user. </br>There should not be any missing values. |

| date | The date the health data was recorded or the supplement was taken, in date format. </br>There should not be any missing values. |

| email | Contact email of the user. </br>There should not be any missing values. |

| user\_age\_group | The age group of the user, one of: 'Under 18', '18-25', '26-35', '36-45', '46-55', '56-65', 'Over 65' or 'Unknown' where the age is missing.|

| experiment\_name | Name of the experiment associated with the supplement usage. </br>Missing values for users that have user health data only is permitted. |

| supplement\_name | The name of the supplement taken on that day. Multiple entries are permitted. </br>Days without supplement intake should be encoded as 'No intake'. |

| dosage\_grams | The dosage of the supplement taken in grams. Where the dosage is recorded in mg it should be converted by division by 1000.</br>Missing values for days without supplement intake are permitted. |

| is\_placebo | Indicator if the supplement was a placebo (true/false). </br>Missing values for days without supplement intake are permitted. |

| average\_heart\_rate | Average heart rate as recorded by the wearable device. </br>Missing values are permitted. |

| average\_glucose | Average glucose levels as recorded on the wearable device. </br>Missing values are permitted. |

| sleep\_hours | Total sleep in hours for the night preceding the current day’s log. </br>Missing values are permitted. |

| activity\_level | Activity level score between 0-100. </br>Missing values are permitted. |



