

Data Engineering: Data Validation

Task 1

Create a folder 'input' and a folder 'output' on your laptop and create excel files from the datasets of last week (Titanic, Iris, Wine and Crime). Save these excel files in the input folder.

Task 2

Create a general Reader class that reads all the files in the input folder.

Task 3

Create a general Writer class that writes the end result to the output folder.

Task 4

Create four specific Validation class's (for each dataset specific) that validates each column. Pay attention to all the types of validation from the course. This validation class's input is the output of the Reader Class. You should also log everything such that you know where the errors occur.

Task 5

Create four specific Processor class's (for each dataset specific) that implements the transformations from the previous exercise sessions.

You may skip the processing step of 'Calculate the crime rate per 100 000 people in each ward'

Task 6

Create a main function where you link all your steps together.

Task 7

Test your code by adding the files from Toledo into your input folder and run the code.

Tell me what is wrong with each file, where the errors occur.

Task 8 (bonus)

Create a Back up validation class to validate the columns after the processing step.