Remote Data Task

Introduction

An e-commerce shop would like to onboard new suppliers efficiently. To enable the onboarding process, the customer needs us to integrate product data from suppliers in various formats and styles into the pre-defined data structure of their e-commerce shop application.

Input files

- Supplier: supplier_car.json
- Target: Target Data.xlsx

Tasks

Your goal is to transform the supplier data so that it could be directly loaded into the target dataset without any other changes. For each step, you should first profile the data to understand what you can do for the customer, then implement a few selected functions (you can keep it lightweight) and tell the customer what you can also potentially do for them, using examples to illustrate.

1) Step Pre-processing

Here you need to transform the supplier data to achieve the same granularity as the target data.

(Hint: how many rows per product do you have in the target data?).
Be aware of character encodings when processing the data.

2) Step Normalisation

Normalisation is required in case an attribute value is different but actually is the same (different spelling, language, different unit used etc.).

Please normalise at least 2 attributes, and describe which normalisations are required for the other attributes including examples.

- Input: pre-processed data
- Output: normalised supplier data

3) Step Integration

Data Integration is to transform the supplier data with a specific data schema into a new dataset with target data schema, such as to:

- keep any attributes that you can map to the target schema
- discard attributes not mapped to the target schema
- keep the number of records as unchanged
- Input: normalised supplier data
- Output: integrated supplier data

Deliverables

- An Excel/LibreOffice spreadsheet (**no csv, no txt**) with 3 tabs showing the results of each step above (i.e., pre-processing/normalisation/integration)
- A script (R/Python/SQL/etc.) that can be executed to provide the above Excel file
- A customer presentation (PowerPoint or similar) to describe the above processing. Assume that the audience is the customer onboarding manager – someone with business knowledge, and medium technical knowledge. This presentation should include at least:
    - key facts of input / output, e.g., # attributes, # records
    - summary of changes you made to the input data
    - summary of changes you can potentially make to the input data, with examples of potentially affected products
    - take-away message and actions to take for the customer

Note that the customer is not interested in data analysis (correlation, histograms, distributions of attributes, etc.). Please only provide relevant information for the business case.