CHALLENGE HINTS

SECOND CHALLENGE

# USE PYTHON TO CLEANSE AND RATIONALISE DATA

The flowchart illustrates one way of answering the challenge. You are not compelled to use this method.

To cater for the language issues three lists are created, e.g. **lookupFrench = ["zero","un","deux","trois",…]** … etc, to twenty (vingt)

Lists for English and German are created in a similar fashion. These lists may then be checked to see if a word is in the list, e.g.

**if (item in lookupFrench): value = lookupFrench.index(item)**

The test data is placed into a “list of lists”:

**data = [["16", "10", "8", "3", "7"], ["8", "09", "19", "20", "4"],**

**["Sechs", "Acht", "Sechzehn", "Funf", "null"],**



**["1", "30", "2", "5", "7"],**

**["Vierzehn", "Eins", "zwei", "Neun", "Drei"],**

**["six", "neuf", "seize", "zero",""],**

**["fourteen", "Eleven", "Forteen", "eight", "Twenty"],**

**["Douze", "Onze", "Huit", "Quinze", "Sept"],**

**["018", "09", "09", "022", "04"],**

**["un", "trois", "quatorze", "dix-huit", "vingt"],**

**["Five", "Three", "Nineteen", "Twenty", "zero"],**

**["einundzwanzig", "Vierzehn", "Eins", "zwei","Vier"]]**

The method illustrated uses two loops:

1. The first loop is an outer loop iterating around each row of data
2. The second loop is created inside the first loop, iterating around each item of data in each row

As each row is processed, the inner loop checks each item in different ways:

1. The item is checked to see if it is numeric (Python string function **isdigit()**).
   1. If it is numeric then the value can be cast to an integer (Python: **(int)str**) and the value checked to see if it is in the range 0-20. If it is it may be stored in a temporary list until all items in the row have been processed.
   2. If it is not in range it is invalid and the whole row is stored in a list of rejected rows and the remaining items (if any) are skipped.
   3. If it is not numeric it needs to be checked to see if it is contained in one of the lists referred to above.
2. Check for the existence of the item in each of the lookup lists until either:
   1. The item is found and the value may be determined as above.
   2. All lists have been checked and the item has not been found. This means that the value is not a number (as text) or is out of range (“twenty-two” is a valid number but is not in the list as it is out of range.)

As each row is completed, if successful it is stored in a “valid” list or in a “rejected” list, if not. The program exits when all rows have been checked.

An alternative method could omit checking for a numeric value and use another list of the numbers as string items, e.g.

**lookupNumbers = ["1","2","3", ...] etc.**

The flowchart overleaf is for the first method.

2

Start

Get data row

Check data item

Is it numeric

?

Get first

/

next data item

In range

?

Check in lookup

for English

Yes

No

Check in lookup

for French

Check in lookup

for German

Not

there

Not

there

OK so store temporarily

Store row

In rejected list

Not

there

Finished with

data row

?

**This is the outer loop**

**.**

**It iterates once for**

**each row of data**

**This is the inner loop**

**.**

**This iterates once for**

**each item in a row**

**.**

Store row

In accepted list

No

Yes

**Any occurrence of an invalid**

**item means that no more**

**checking needs to be**

**performed**

**.**

**Store the row**

**data and move on to the next**

**row**

**.**

Yes

No

More rows

?

Yes

End

No

3