Framework: ETL Data Pipeline: Extract, Transform, Load

Languages: Python, pandas, ...

Input: Raw data file

Output: Updated data template - CSV

Solution: have consistent permanent titles names

Possible problems:

- loading data in the correct cells in an excel spreadsheet
- Sample ID's change depending on the sample

Random Thoughts:

- Have a button in the excel sheet which runs the program, could have input values which we might need to do some of the equation such as MDL etc.

----------

Titles for every run:

- ICV
- ICB
- MDL
- CCV#... -> it will have CCV1, CCV2 ... CCV#, so we need to figure this out
- CCB#... -> it will have CCB1, CCB2 ... CCB#, so we need to figure this out


Other QC Sample names:

- QCS
- QCB
- MS
- MSD

----------

Rows needed:

- Sample ID - G
- Mean – I
- PPM - L

----------

**Info needed for Calibration: -> doesn't run all the time, graph is using to check if calibration is correct, so if R^2 is close to 1**

- Curve value -> found in between bracket [TOC] TOC 20 ppm CAL [0.4]
    - In the calibration sheet, are these set in the template? -> YES
- Adjusted ABS -> values given to us depending if we do a calibration test
- Average -> Average of the two runs, we could use an equation in python
- %RPD -> ((absolute value of run1 – run2 )/ Average) * 100

Notes:

- %RPD < 10
- Will need to output graph

----------

**Info needed for Quality Control:**

Sample ID's needed:

- MDL -> is it always 0.2?
- QCS -> is it always 18?
- CCV# -> # changes from 1 till …
- QCB
- CCB# -> # changes from 1 till…

Mean PPM C:

- Average of the 2 runs of MDL, QCS, CCV1, CCV2, CCV#... etc.

%R:

- (Mean PPM C / Value) * 100
    - Value depends on title, for example MDL(0.2), has 02, while CCV is 10

%RPD:

- %RPD -> ((absolute value of run1 – run2 )/ Mean PPM C) * 100

Average:

- Average of All QCB and CCB values.

Notes:

- %R MDL-45-145
- %R 90 – 110? Mario said to change it to this