Data Cleaning Exercise

Use Python to download a raw data file and clean it.

Statement of the problem

- Use python ‘requests’ package to get the file from here https://raw.githubusercontent.com/victorbrub/data-engineering-class/refs/heads/main/pre-post_processing/exercise.csv
- Check the file data to fast check. Read it with Pandas.
    - How many rows do we have?
    - Is there any sensible information?
    - What kind of problems can we have regarding the nature of this data?
- Clean it.
    - Define the rules we need to clean the data.
    - Implement a Python Script that cleans the data and analyzes the clean process.
- Output should be a new file with cleaned data.
- Create a test on the raw data and on the cleaned data, for each one of the dimensions. The output should be a percentage: rows that meet the test requirement / total rows. Compare the results of each file.

Data Quality Dimensions

Dimension	**Definition**	**Example**	**Impact**
Accuracy	Data represents reality correctly	Age should be realistic (not 999 years)	Wrong decisions
Completeness	All required data is present	Every customer must have an email	Missing insights
Consistency	Data is uniform across systems	"USA" vs "United States"	Failed joins
Validity	Data conforms to required format	Email must have @ symbol	System errors
Uniqueness	No duplicate records exist	One OrderID per order	Inflated metrics
Timeliness	Data is current and available	Latest customer address	Outdated insights