DATA CLEANING EXERCISE

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

USJ
INGENIERIA DE DATOS CURSO 2025-2026

FRANCISCO JAVIER LIARTE NASARRE

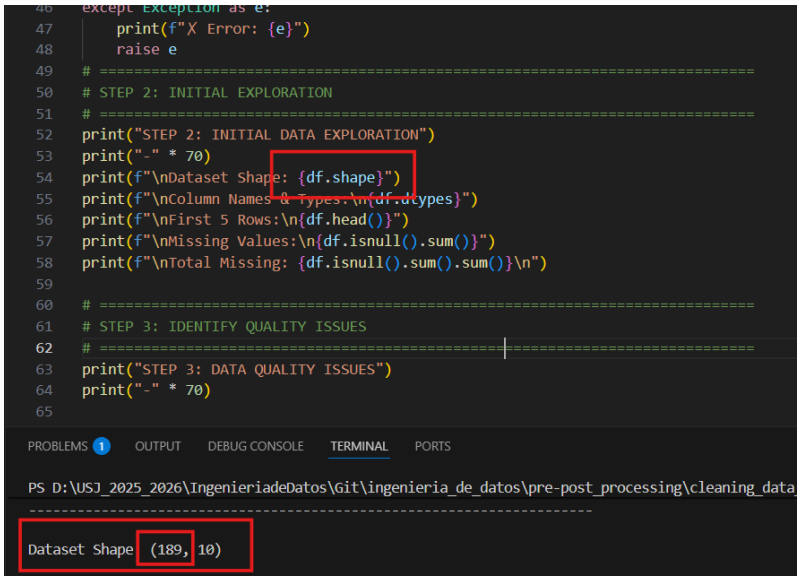
Statement of the problem

• Use python ‘requests’ package to get the file from here

https://raw.githubusercontent.com/victorbrub/data-engineeringclass/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?



We have 189 rows

• Is there any sensible information?

Yes, I mean, with the CustomerName, Phone and Age, you can easily find that person nowadays (if he does not use some type of alias or has other security options)

• What kind of problems can we have regarding the nature of this data?

Looking at the data we can see is from a shop or some warehouse. This data is very important, because is crucial to keep the orders of the clients. This data can be used to solve issues and to keep records of the orders of the clients and create analyses to improve the shop’s performance.

• Clean it.

In the step of Python Script.

• Define the rules we need to clean the data.

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Timeliness	Data is current and available	Latest customer address	Outdated insights

Following this table, we will do the following for the columns:

OrderID:
Missing -> Drop It. Is the primary key so if there is an error is the easiest way to solve it
Duplicate -> Drop em.

CustomerName:
Missing -> Impute unknown
Validate the format -> Name Surname, if is only changing the letters, reformate it.

Email:
Missing -> Drop It
Validate the format -> juan@email.com if is only changing the letters, reformate it.

Phone:
Missing -> Drop It
Validate the format -> 555-XXXX if incorrect drop it

Country:
Missing -> Impute unknown
Validate the format -> CAPITAL if is only changing the letters, reformate it.

OrderDate:
Missing -> Drop It
Validate the format -> min-> 1/1/2023 max -> 31/12/2023 if incorrect, drop it

Quantity:
Missing -> Drop It
Validate the format -> INT, if no drop it

Price:
Missing -> Drop It
Validate the format -> INT, if no drop it

CustomerAge:
Missing -> Impute media()
Validate the format -> between 18-99 if incorrect drop it

OrderStatus:
Missing -> Drop It

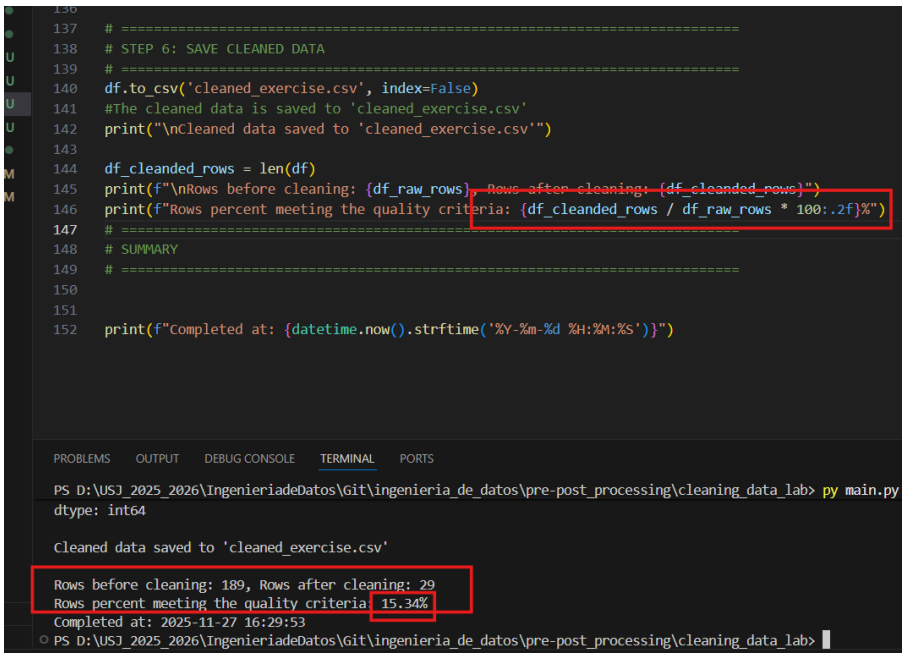
• Implement a Python Script that cleans the data and analyzes the clean process.

Check the python files in my github repository: https://github.com/liliarte-1/ingenieria_de_datos#

• Output should be a new file with cleaned data.

In step 6, in the cleaned_exercise.csv.

• 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.



As we can see, this data retrieval needs to be improved.