Challenge 2 is HERE!!

Your real step to transform the raw data to ML/AI ready data with :

Data Preprocessing !

Objective :

Transform The house pricings dataset from messy, unclear data to clean structured format ready for machine learning, using the concepts taught in the 3rd workshop, and split it to training and testing.

Dataset:

We'll use the House Prices competition dataset :

Here: https://www.kaggle.com/competitions/home-data-for-ml-course/data

The files that you need are:

Data_description.txt which provides description of all the features

Train.csv that contains the data set

Challenge tasks explanation :

Load and analyze your data :

- Import the important libraries
- Load your data set , and display necessary stats
- Display categorical features
- Display which features have missing values
- Plot features

Time to clean the data using Preprocessing and Feature engineering :

- Drop unnecessary features
- Create new features if possible
- Handle missing values using the right techniques (filling with stats/ time series)
- Encode categorical data using Label-Encoding or One-Hot Encoding (depends on which case is better)
- Use data scaling if needed

Define your main X features and target Y

Train-Test Split :

- Split your dataset after finishing cleaning into two sets one for training and one for testing
- Explain why we split

Deliverables :

A notebook containing :

- All code for the tasks specified above
- Comments explaining each step
- Summary (10 lines max) explaining the decisions you made to clean the dataset

Ex : when encoding categorical data, why did you use One-Hot encoding instead of label encoding?

- Compare the Before cleaning and After cleaning data, what do you conclude?
- Answer the question : *"What would happen if we trained the model without these steps?"*

Always remember that , The data preprocessing is the most important step for making a successful AI project , understanding your data and the well organization saves you time in crucial moments.

GOOD LUCK!