Lab Assignment: Detecting Outliers in House Prices

Dataset:

Ames Housing Dataset
Link: https://www.kaggle.com/datasets/prevek18/ames-housing-dataset

Description:
A detailed dataset of house prices in Ames, Iowa, with **82 features** covering building size, construction year, materials, garage, basement, etc.

Lab Tasks:

1. Download and load the **Ames Housing dataset** using Pandas.
2. Select a numerical feature (e.g., **SalePrice**) and calculate Z-Scores.
    - Identify values where $|\mathbf{Z}| > \mathbf{3}$.
3. Apply the **IQR method** to the same feature and detect outliers.
4. Apply **Isolation Forest** on *all* numerical features of the dataset.
5. Compute and compare:
    - Number of Z-Score outliers
    - Number of IQR outliers
    - Number of Isolation Forest anomalies
6. Visualize:
    - A boxplot of the selected feature
    - A scatter plot (e.g., **GrLivArea vs SalePrice**) highlighting Isolation Forest anomalies
7. Write a short comparison:
    - Which method detected the most anomalies?
    - Which method seems the most reasonable?

Exercises:

1. Why might the Z-Score method fail when the dataset contains strongly skewed price distributions?
2. Why is the IQR method more resistant to extreme values? Explain with an example.
3. In Isolation Forest, what happens if you set `contamination=0.10` in a dataset where outliers are actually rare?
4. Which method (Z-Score, IQR, Isolation Forest) would be best for real-estate price anomaly detection? Justify your answer.