# In-Class Practices

1. Environment setup:
   1. Python (runtime env);
   2. Pip or Conda (package management);
   3. JupyterLab (Notebook kernel);
   4. VSCode (IDE);
   5. Airflow (this required linux OS – for windows users it’s recommended to have WSL2 installed);
   6. Dependencies: great\_expectations, pandas, numpy, sckit-learn, …
2. EDA & Cleaning data with Pandas: Kaggle datasets & practices
   1. Lab: data\_prep.ipynb
   2. [Learn Data Cleaning | Kaggle](https://www.kaggle.com/learn/data-cleaning)
   3. [Cleaning Data using Pandas](https://www.kaggle.com/code/dograayushh/cleaning-data-using-pandas)
   4. [Data ScienceTutorial for Beginners](https://www.kaggle.com/code/kanncaa1/data-sciencetutorial-for-beginners#3.CLEANING-DATA)
3. Explore fundamentals of Great Expectations (GX)
   1. Documentations: https://greatexpectations.io/
   2. Lab: gx.ipynb
4. Configure Airflow DAG to concatenate multiple steps: pulling data, cleaning, ensure data quality, training model from cleaned datasets: Instruction-Led Training
   1. Documentation: https://airflow.apache.org/
   2. Lab: dag-sample.py