Homework Sheet — Stage 6: Data Preprocessing

Assignment

You will extend today’s in-class work by writing modular cleaning functions, applying them to the provided raw dataset, and documenting your assumptions.

Instructions

1. Create functions in `src/cleaning.py`:
   - `fill_missing_median()`
   - `drop_missing()`
   - `normalize_data()`
2. Use the provided raw dataset in `/data/raw/`.
3. In a Jupyter notebook:
   - Load dataset
   - Apply cleaning functions
   - Save cleaned dataset to `/data/processed/`
   - Compare original vs cleaned data
   - Document all assumptions clearly
4. Update README with a section on cleaning strategy.

Explicit Chain

In the lecture, we learned how to **fill, drop, and scale** data systematically. Now, you will **adapt these methods into reusable functions and apply them to your dataset**.

Submission

- `src/cleaning.py`
- Preprocessing notebook
- Updated README
- Saved processed dataset

Grading Rubric

- **Correctness (40%)**: Functions work as expected
- **Documentation (20%)**: Clear assumptions, docstrings, README updates
- **Reproducibility (20%)**: Clean dataset saved correctly, code modular
- **Reflection (20%)**: Notebook comparisons and notes on tradeoffs

Example Expectations

- Code functions with docstrings
- Notebook demonstrating transformations
- Clear README documentation
- Saved cleaned dataset